2019 AMIA Board of Directors

Officers

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

Patricia C. Dykes, PhD, RN, FAAN, FACMI
Brigham and Women’s Hospital
Chair-elect

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

Theresa A. Cullen, MD
Regenstrief Institute
Secretary

Directors

Tiffany J. Bright, PhD
Washington University School of Medicine in St. Louis

Wendy W. Chapman, PhD, FACMI
University of Utah

James J. Cimino, MD
University of Alabama at Birmingham Informatics Institute School of Medicine

David Dorr, MD, MS, FACMI
Oregon Health & Science University

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

Wanda Pratt, PhD, FACMI
University of Washington

Gretchen Purcell Jackson, MD, PhD
IBM Watson Health/Vanderbilt University Medical Center

S. Trent Rosenbloom, MD, MPH, FACMI
Vanderbilt University Medical Center

Sarah (Collins) Rossetti, RN, PhD
Partners Healthcare System/Brigham and Women’s Hospital/Harvard Medical School

Neil Sarkar, PhD, FACMI
Brown University

Jessica Tenenbaum, PhD
Duke Translational Medicine Institute

Adam Wright, PhD
Vanderbilt University Medical Center

**Ex-Officio Board Members**

William M. Tierney, MD, FACMI
American College of Medical Informatics President
University of Texas at Austin

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

Stephen Johnson, PhD, FACMI
Academic Forum Executive Committee Chair
Weill Cornell Medicine

Albert M. Lai, PhD
Working Group Steering Committee Representative
Washington University

Shauna M. Overgaard, MHI, PhD Candidate
Student Working Group Representative
University of Minnesota

Frank Naeymi-Rad, PhD, MBA, FACMI
Informatics Partnership Council Chair
Intelligent Medical Objects
AMIA 2019 Annual Symposium Scientific Program Committee

Chair
Elmer V. Bernstsam, MD, FACMI
University of Texas Health Science Center

Vice Chairs
Dina Demner-Fushman, MD, FACMI
National Library of Medicine

S. Trent Rosenbloom, MD, MPH, FACMI
Vanderbilt University Medical Center

Laura K. Wiley, PhD
University of Colorado

Adam Wright, PhD, FACMI
Vanderbilt University Medical Center

Members
Joanna Abraham, PhD
Institute for Informatics at Washington University in St. Louis, School of Medicine

Alexander V. Alekseyenko, PhD
Medical University of South Carolina

Allison Beck McCoy, PhD
Vanderbilt University Medical Center

Jiang Bian, PhD
University of Florida

Mary Regina Boland, PhD
University of Pennsylvania

Jonathan P. Bona, PhD
University of Arkansas for Medical Sciences

Suzanne A. Boren, PhD
University of Missouri

Andrew D. Boyd, MD
University of Illinois at Chicago
Juliana Brixey, PhD, MPH, RN
University of Texas Health Science Center

Brian Thomas Bucher, MD
University of Utah/Primary Children's Hospital

Thomas R. Campion, PhD
Weill Cornell Medical College

Pamela Charney, RD, PhD
University of North Georgia

You Chen, PhD
Vanderbilt University

Catherine Craven, PhD
Icahn School of Medicine at Mount Sinai

Robert Cronin, MD
Vanderbilt University Medical Center

Berry de Bruijn, PhD
National Research Council, Digital Technologies Research Centre

Judith W. Dexheimer, PhD
Cincinnati Children’s Hospital

Hossein Estiri, PhD
Harvard Medical School

Jungwei Fan, PhD
Mayo Clinic

Susan H. Fenton, PhD
UT School of Biomedical Informatics

Adela Grando, PhD
Arizona State University

Robert A. Jenders, MD, FACP, FACMI
Charles Drew University/UCLA

Thomas Kannampallil, PhD
Washington University School of Medicine
Ramakanth Kavuluru, PhD
University of Kentucky, College of Medicine

Shameer Khader, PhD
AstraZeneca

Saif Khairat, PhD
University of North Carolina

Eric Kirkendall, MD, MBI
Wake Forest Baptist Medical Center

Mei Liu, PhD
University of Kansas Medical Center

Hongfang Liu, PhD
Mayo Clinic

Yuan Luo, PhD
Northwestern University

Bradley A. Malin, PhD
Vanderbilt University Medical Center

Ruth M. Masterson Creber, PhD, MSc, RN
Weill Cornell Medicine

Stephane M. Meystre, MD, PhD
Medical University of South Carolina

Jason M. Mitchell, MD, MS
Cerner Corporation

Stuart J. Nelson, MD, FACP, FACMI
University of New Mexico School of Medicine

Aurelie Neveol, PhD
LIMSI – CNRS

Omolola I. Ogunyemi, PhD
Charles Drew University of Medicine and Science

Carolyn Petersen, MS, MBI
Mayo Clinic

**Andrew B. Phillips**, PhD, RN  
MGH Institute of Health Professions, Massachusetts General Hospital

**Ronald J. Piscotty**, PhD, RN-BC  
University of Maryland

**Lee Pyles**, MD  
West Virginia University

**Luke Rasmussen**  
Northwestern University

**Janise Richards**, MS, MPH, PhD  
Centers for Disease Control & Prevention

**Satya S. Sahoo**, PhD  
Case Western Reserve University

**Hojjat Salmasian**, MD, MPH, PhD  
Brigham and Women's Hospital

**David F. Schneider**, MD  
University of Wisconsin

**Murat Sincan**, MD  
Sanford Health

**Neil R. Smalheiser**, MD  
University of Illinois at Chicago

**Sandeep Soman**, MD  
Henry Ford Health System

**Joshua H. Tamayo-Sarver**, MD, PhD  
CEP America

**Ozlem Uzuner**, PhD  
George Mason University

**Theresa Walunas**, PhD  
Northwestern University
Amy Y. Wang, MD, MBI, FAAFP
University of Alabama at Birmingham Informatics Institute School of Medicine

Keith F. Woeltje, MD, PhD
BJC HealthCare

Po-Yin Yen, PhD, RN
Washington University in St Louis

Kai Zheng, PhD
University of California, Irvine
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aarts, Jos</td>
<td>Ahmad, Faraz</td>
<td>Amato, Mary</td>
</tr>
<tr>
<td>Abbott, Amy</td>
<td>Ahmad, Suzan</td>
<td>Ambert, Kyle</td>
</tr>
<tr>
<td>Abdelrahman, Samir</td>
<td>Ahmed, Adil</td>
<td>Ambite, Jose-Luis</td>
</tr>
<tr>
<td>Abedtash, Hamed</td>
<td>Ahmed, Zeeshan</td>
<td>Ambrosino, Richard</td>
</tr>
<tr>
<td>Abel, Kingsley</td>
<td>Airan-Javia, Subha</td>
<td>Ameri, Kimia</td>
</tr>
<tr>
<td>Abend, Aaron</td>
<td>Ajjarapu, Samuel</td>
<td>Amiri, Hadi</td>
</tr>
<tr>
<td>Abeysinghe, Rashmie</td>
<td>Akter, Sadia</td>
<td>Amith, Muhamamd</td>
</tr>
<tr>
<td>Abhyankar, Swapna</td>
<td>Alafaireet, Patricia</td>
<td>Amland, Robert</td>
</tr>
<tr>
<td>Abraham, Joanna</td>
<td>Al-alawneh, Saja</td>
<td>Ammenwerth, Elske</td>
</tr>
<tr>
<td>Abrams, Meredith</td>
<td>Albers, David</td>
<td>Anand, Vibha</td>
</tr>
<tr>
<td>Abujarad, Fuad</td>
<td>Alexander, Nicholas</td>
<td>Anders, Shilo</td>
</tr>
<tr>
<td>Aczon, Melissa</td>
<td>Alhuwail, Dari</td>
<td>Anderson, Nick</td>
</tr>
<tr>
<td>Adam, Terrence</td>
<td>Alipanah, Neda</td>
<td>Anderson, William</td>
</tr>
<tr>
<td>Adapa, Karthik</td>
<td>Aljabban, Imad</td>
<td>Andrade, Jorge</td>
</tr>
<tr>
<td>Adegoke, Oluwasegun</td>
<td>Aljabban, Jihad</td>
<td>Andrews, Steven</td>
</tr>
<tr>
<td>Adekkanattu, Prakash</td>
<td>Allen, Bryce</td>
<td>Anjomshoa, Hamideh</td>
</tr>
<tr>
<td>Adibuzzaman, Mohammad</td>
<td>Allen, Katie</td>
<td>Antani, Sameer</td>
</tr>
<tr>
<td>Adrian, Bonnie</td>
<td>Alonso, William</td>
<td>Anton, Bonnie</td>
</tr>
<tr>
<td>Adrianto, Indra</td>
<td>Alpi, Kristine</td>
<td>Apathy, Nate</td>
</tr>
<tr>
<td>Afshar, Majid</td>
<td>Al-qaqaa, Yasir</td>
<td>Apostolova, Emilia</td>
</tr>
<tr>
<td>Agarwal, Kanupriya</td>
<td>Alsentzer, Emily</td>
<td>Araj, Ellen</td>
</tr>
<tr>
<td>Agresta, Thomas</td>
<td>Altamore, Rita</td>
<td>Arcan, Chrisa</td>
</tr>
<tr>
<td>Agu, Nkechinyere Nneka</td>
<td>Amadi, Tochi</td>
<td>Ariosto, Deborah</td>
</tr>
<tr>
<td>Name</td>
<td>First Name</td>
<td>Last Name</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Arruda-Olson, Adelaide</td>
<td>Bangar, Suhasini</td>
<td>Bennett, Tellen</td>
</tr>
<tr>
<td>Artz, Nathan</td>
<td>Banger, Alison</td>
<td>Ben-Omran, Munder</td>
</tr>
<tr>
<td>Asagbra, Oghale</td>
<td>Bao, Riyue</td>
<td>Bergamasco, Leila</td>
</tr>
<tr>
<td>Asan, Onur</td>
<td>Barnhill, Rick</td>
<td>Bergant, Gaber</td>
</tr>
<tr>
<td>Asaro, Phillip</td>
<td>Barnholtz-Sloan, Jill</td>
<td>Berghout, Joanne</td>
</tr>
<tr>
<td>Ashcraft, Kristine</td>
<td>Barretto, Eluizio</td>
<td>Bergquist, Timothy</td>
</tr>
<tr>
<td>Ashtari, Sadaf</td>
<td>Baskett, William</td>
<td>Bernard, Tamara</td>
</tr>
<tr>
<td>Aslani, Azam</td>
<td>Bass, Ellen</td>
<td>Bernstam, Elmer</td>
</tr>
<tr>
<td>Aucoin, Jennifer</td>
<td>Bass, Michael</td>
<td>Berrios, Daniel</td>
</tr>
<tr>
<td>Averitt, Amelia</td>
<td>Baxter, Sally</td>
<td>Berry, Donna</td>
</tr>
<tr>
<td>Avramovic, Sanja</td>
<td>Baysari, Melissa</td>
<td>Berry, Shamsi</td>
</tr>
<tr>
<td>Avrunin, George</td>
<td>Bean, Carol</td>
<td>Besiso, Ali</td>
</tr>
<tr>
<td>Aziz, Md Momin</td>
<td>Bear Don't Walk IV, Oliver</td>
<td>Bettencourt-Silva, Joao</td>
</tr>
<tr>
<td>Baek, Jeong-Heum</td>
<td>Beebe, Russ</td>
<td>Beukenhorst, Anna</td>
</tr>
<tr>
<td>Bagley, Steven</td>
<td>Beheshti, Rahmatollah</td>
<td>Bhattacharya, Moumita</td>
</tr>
<tr>
<td>Bakal, Mehmet</td>
<td>Behiri, Eytan</td>
<td>Bhavnani, Suresh</td>
</tr>
<tr>
<td>Bako, Abdulaziz</td>
<td>Beidas, Sary</td>
<td>Bhuvaneswar, Krithika</td>
</tr>
<tr>
<td>Balatsoukas, Panagiotis</td>
<td>Bellam, Shashi</td>
<td>Bian, Jiang</td>
</tr>
<tr>
<td>Baldwin, Jessica</td>
<td>Bellazzi, Riccardo</td>
<td>Binkheder, Samar</td>
</tr>
<tr>
<td>Baldwin, Tyler</td>
<td>Belouali, Anas</td>
<td>Blais, Rebecca</td>
</tr>
<tr>
<td>Balyan, Renu</td>
<td>Ben Miled, Zina</td>
<td>Blake, Catherine</td>
</tr>
<tr>
<td>Banach, Mary</td>
<td>Ben Said, Mohamed</td>
<td>Blauffeux, Brian</td>
</tr>
<tr>
<td>Banerjee, Imon</td>
<td>Benda, Natalie</td>
<td>Blecker, Saul</td>
</tr>
<tr>
<td>Banerjee, Tanvi</td>
<td>Bennett, Kristin</td>
<td>Blondon, Katherine</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Chase, Herbert</td>
<td>Cheng, Eric</td>
<td>Chung, Jane</td>
</tr>
<tr>
<td>Che, Zhengping</td>
<td>Cheng, Hao-Yuan</td>
<td>Cimino, James</td>
</tr>
<tr>
<td>Chen, Annie</td>
<td>Cheng, Yan</td>
<td>Clark, Christine</td>
</tr>
<tr>
<td>Chen, Chen</td>
<td>Chester, Kelley</td>
<td>Clarkson, Melissa</td>
</tr>
<tr>
<td>Chen, Ching-Hua</td>
<td>Chhabra, Divya</td>
<td>Clee, Mark</td>
</tr>
<tr>
<td>Chen, David</td>
<td>Chhikara, Jatin</td>
<td>Clyde, Stephen</td>
</tr>
<tr>
<td>Chen, Jake</td>
<td>Chi, Nai-Ching</td>
<td>Cochrán, Allyson</td>
</tr>
<tr>
<td>Chen, James</td>
<td>Chiaravalloti, Maria</td>
<td>Cohen, Dasha</td>
</tr>
<tr>
<td>Chen, Jessica</td>
<td>Chih, Ming-Yuan</td>
<td>Cohen, Trevor</td>
</tr>
<tr>
<td>Chen, Ji</td>
<td>Chin, Chieh-Li</td>
<td>Colicchio, Tiago</td>
</tr>
<tr>
<td>Chen, Jinying</td>
<td>Chin, Jessie</td>
<td>Colletta, Sarah</td>
</tr>
<tr>
<td>Chen, Jonathan</td>
<td>Chiu, Po-Hsiang</td>
<td>Collins Rossetti, Sarah</td>
</tr>
<tr>
<td>Chen, Jun</td>
<td>Cho, Hwayoung</td>
<td>Comer, Karen</td>
</tr>
<tr>
<td>Chen, Long</td>
<td>Cho, Insook</td>
<td>Conboy, Heather</td>
</tr>
<tr>
<td>Chen, Lujie</td>
<td>Cho, Seoyoon</td>
<td>Conic, Rosalynn</td>
</tr>
<tr>
<td>Chen, Luoxin</td>
<td>Choi, Jeeyae</td>
<td>Connolly, Elizabeth</td>
</tr>
<tr>
<td>Chen, Minxing</td>
<td>Chokshi, Sara</td>
<td>Connor, MaryAnn</td>
</tr>
<tr>
<td>Chen, Qingcai</td>
<td>Chou, Chia-Ching</td>
<td>Conway, Mike</td>
</tr>
<tr>
<td>Chen, Qingxiao</td>
<td>Chou, Jingyuan</td>
<td>Corcorán, Kelsey</td>
</tr>
<tr>
<td>Chen, Qingyu</td>
<td>Choudhury, Olivia</td>
<td>Corley, Sarah</td>
</tr>
<tr>
<td>Chen, Tianlong</td>
<td>Chu, Chia-Hui</td>
<td>Cossul, Sandra</td>
</tr>
<tr>
<td>Chen, Tingting</td>
<td>Chu, Jiangfeng</td>
<td>Crane, Heidi</td>
</tr>
<tr>
<td>Cheng, Alex</td>
<td>Chun, Soon Ae</td>
<td>Craven, Catherine</td>
</tr>
<tr>
<td>Cheng, Daryl</td>
<td>Chung, Arlene</td>
<td>Cronkite, David</td>
</tr>
<tr>
<td>Crooks, John</td>
<td>de Bruijn, Berry</td>
<td>Do, Nhue</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Crotty, Tina</td>
<td>De Castro, Alexandre</td>
<td>Doan, Son</td>
</tr>
<tr>
<td>Crowley Jacobson, Rebecca</td>
<td>de Oliveira Sales, Ana</td>
<td>Doebbeling, Bradley</td>
</tr>
<tr>
<td>Cui, Licong</td>
<td>Dearth, Shandy</td>
<td>Dolin, Gay</td>
</tr>
<tr>
<td>Cui, Zhicheng</td>
<td>Del Zotto, Marzia</td>
<td>Dong, Xiao</td>
</tr>
<tr>
<td>Culbreath, Chandrea</td>
<td>Deleger, Louise</td>
<td>Dorr, David</td>
</tr>
<tr>
<td>Cummins, Mollie</td>
<td>Della Mea, Vincenzo</td>
<td>Doshi, Riddhi</td>
</tr>
<tr>
<td>Curcin, Vasa</td>
<td>Delman, Bradley</td>
<td>Douthit, Brian</td>
</tr>
<tr>
<td>Curtis, Dorothy</td>
<td>Demiris, George</td>
<td>Downing, Lance</td>
</tr>
<tr>
<td>Cusack, Caitlin</td>
<td>Demner-Fushman, Dina</td>
<td>Downs, Sarah</td>
</tr>
<tr>
<td>Cyran, Marek</td>
<td>Deng, Yu</td>
<td>Doyon, Katherine</td>
</tr>
<tr>
<td>Dahm, Maria</td>
<td>Deshmukh, Vikrant</td>
<td>Drake, Connor</td>
</tr>
<tr>
<td>Dai, Peng</td>
<td>Devarakonda, Murthy</td>
<td>Driscoll, Patricia</td>
</tr>
<tr>
<td>Dalcour, Cortney</td>
<td>Dewan, Maya</td>
<td>Du, Fan</td>
</tr>
<tr>
<td>Daphtary, Kshama</td>
<td>Dexheimer, Judith</td>
<td>Du, Shimeng</td>
</tr>
<tr>
<td>Dasgupta, Dipanwita</td>
<td>Dhopeshwarkar, Neil</td>
<td>Du, Xinsong</td>
</tr>
<tr>
<td>Dasgupta, Pritika</td>
<td>Diallo, Gayo</td>
<td>Du, Yao</td>
</tr>
<tr>
<td>Datta, Surabhi</td>
<td>Diane, Anna</td>
<td>Dua, Prerna</td>
</tr>
<tr>
<td>Davidson, Patricia</td>
<td>Diaz, Diana</td>
<td>Dubovitskaya, Alevtina</td>
</tr>
<tr>
<td>Davis, Katherine</td>
<td>Diaz-Garelli, Jose-Franck</td>
<td>Duke, Jon</td>
</tr>
<tr>
<td>Davis, Konrad</td>
<td>Diehl, Alexander</td>
<td>Dullabh, Prashila</td>
</tr>
<tr>
<td>Davis, Sharon</td>
<td>Ding, Ying</td>
<td>Dunn, Adam</td>
</tr>
<tr>
<td>Davoudi, Anis</td>
<td>Divita, Guy</td>
<td>Dunn Lopez, Karen</td>
</tr>
<tr>
<td>Dcruz, Jina</td>
<td>Dixon, Brian</td>
<td>Dunnenbernger, Henry</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Durojaiye, Ashimiyu</td>
<td>Farinelli, Fernanda</td>
<td>Franzke, Laura</td>
</tr>
<tr>
<td>Dutra, Ines</td>
<td>Faysel, Mohammad</td>
<td>Freimuth, Robert</td>
</tr>
<tr>
<td>Edgar, Heather</td>
<td>Fazio, Sarina</td>
<td>Fu Jr., Paul</td>
</tr>
<tr>
<td>Edinger, Tracy</td>
<td>Fedorov, Andriy</td>
<td>Fuji, Kevin</td>
</tr>
<tr>
<td>Edmunds, Margo</td>
<td>Feldman, Henry</td>
<td>Fujita, Katsuhide</td>
</tr>
<tr>
<td>Edwards, Johnathan</td>
<td>Feng, Jinghan</td>
<td>Fultz Hollis, Kate</td>
</tr>
<tr>
<td>Eikey, Elizabeth</td>
<td>Feng, Yuanyuan</td>
<td>Fung, Kin Wah</td>
</tr>
<tr>
<td>Eisenberg, Floyd</td>
<td>Fenton, Susan</td>
<td>Furniss, Stephanie</td>
</tr>
<tr>
<td>Eisenstein, Eric</td>
<td>Fernández Breis, Jesualdo</td>
<td>Gabel, Anne</td>
</tr>
<tr>
<td>Eisman, Aaron</td>
<td>Ferraro PhD, Jeffrey</td>
<td>Gabriel, Davera</td>
</tr>
<tr>
<td>EL HIKI, Lahcen</td>
<td>Few, Barbara</td>
<td>Galichi, Leila</td>
</tr>
<tr>
<td>El Saadawi, Gilan</td>
<td>Fillmore, Nathanael</td>
<td>Galimzianova, Alfiia</td>
</tr>
<tr>
<td>Eldredge, Christina</td>
<td>Finan, Sean</td>
<td>Gallego, Blanca</td>
</tr>
<tr>
<td>El-Kareh, Robert</td>
<td>Flegel, Willy</td>
<td>Gallo De Moraes, Alice</td>
</tr>
<tr>
<td>Elkin, Peter</td>
<td>Florez-Arango, Jose</td>
<td>Galorenzo, Jamie</td>
</tr>
<tr>
<td>Ellingson, Benjamin</td>
<td>Flynn, Allen</td>
<td>Galt, Kimberly</td>
</tr>
<tr>
<td>Erdley, William</td>
<td>Follett, Robert</td>
<td>Galvez, Jorge</td>
</tr>
<tr>
<td>Ergai, Awatef</td>
<td>Fong, Allan</td>
<td>Gamache, Roland</td>
</tr>
<tr>
<td>Everson, Jordan</td>
<td>Ford, Elizabeth</td>
<td>Gangadhar, Shruti</td>
</tr>
<tr>
<td>Fackler, Jim</td>
<td>Forlizzi, Jodi</td>
<td>Ganguly, Debasis</td>
</tr>
<tr>
<td>Falconer, Thomas</td>
<td>Formanek, Perry</td>
<td>Ganslandt, Thomas</td>
</tr>
<tr>
<td>Fan, Liyue</td>
<td>Fort, Daniel</td>
<td>Ganta, Kedar</td>
</tr>
<tr>
<td>Farcas, Emilia</td>
<td>Fossa, Alan</td>
<td>Gao, Cheng</td>
</tr>
<tr>
<td>Fareed, Naleef</td>
<td>Foster, Marva</td>
<td>Gao, Grace</td>
</tr>
</tbody>
</table>
Gao, Yi-Bo
Garcia, Ana
Gardner, Bret
Gartrell, Kyungsook
Garvin, Jennifer
Gaudioso, Carmelo
Gazarian, Priscilla
Ge, Yaorong
Geller, James
Geng, Yimin
Georgsson, Mattias
Gephart, Sheila
Gernant, Stephanie
Gershonik, Esteban
Gevaert, Olivier
Gharibi, Gharib
Ghosh, Debashis
Giannaris, Pericles
Giardina, Traber
Gibbings, Rima
Gibson, Richard
Gilbert, Kevin
Gilder, Jason
Giuffrida, Patricia
Gjekmarkaj, Eva
Goel, Alexander
Goetz Mrs., Kerry
Gold, Rachel
Goldspiel, Barry
Goldstein, Mary
Gomes, Alex
Goncalves, Luciana
Goncalves Rezende Macieira, Tamara
Gong, Jen
Gong, Mengchun
Gong, Yang
Gonzalez Martorell, Eduardo
Goodman PhD, Kenneth
Goodwin, Rebecca
Goodwin, Travis
Gopalan, Raj
Gordon, Geoffrey
Gordon, William
Gorham, Tyler
Goudey, Benjamin
Goyal, Dev
Goyal, Nikhil
Grabar, Natalia
Graetz, Ilana
Graham, Camelia
Grammatico-Guilien, Leslie
Grasso, Michael
Graumlich, James
Graybeal, John
Greene, Sarah
Grieve, Graeme
Griffin, Ashley
Grosjean, Julien
Gross, Brian
Grouin, Cyril
Grout, Randall
Grover, Himanshu
Grundmeier, Robert
Gu, Yang
Guidry, Alicia
Gundlapallli, Adi
Gunn, Natalie
Guntuku, Sharath
Guo, Aixia
Guo, Jia-Wen
Guo, Yufan
Gupta, Aditi
Hossain, Md Mahbub  
Hovey, Lauren  
Hribar, Michelle  
Hron, Jonathan  
Hsiao, Allen  
Hsiao, Chun-Ju  
Hsieh, Yi-Ting  
Hsueh, Pei-Yun  
Hu, Lu  
Hu, Pengwei  
Hu, Peter  
Huang, Catherine  
Huang, Ming  
Huang, Yan  
Huang, Yang  
Hudak, Christine  
Huffman, Jennifer  
Huguet, Nathalie  
HUI, Chi Ching  
Hulse, Nathan  
Hunter-Zinck, Haley  
Huser, Vojtech  
Hussain, Mustafa  
Hylock, Ray  
Hyun, Sookyung  
Ibañez, Carolina  
Ibrahim, Omar  
Imler, Timothy  
Innokenteva, Yulia  
Introne PhD, Josh  
Iott, Bradley  
Isaac, Samuel  
Israni MS, Juhi  
Iturrate, Eduardo  
Jabbari, Fattaneh  
Jackson, Gretchen  
Jacobs, Maia  
Jacobs, Robin  
Jacobs, Tom  
Jadhav, Ashutosh  
Jagannath, Swathi  
Jagannathan, Vasudevan  
Jamieson, Trevor  
Jang, Jeehoon  
Jayakumaran, Gowtham  
Jayapandian, Catherine  
Jeffery, Alvin  
Jenders, Robert  
Jenkins, Melinda  
Jethani, Neil  
Ji, Xiaonan  
Ji, Zongcheng  
Jiang, Dehuan  
Jiang, Silis  
Jiang, Xiaqian  
Jiang, Yun  
Jin, Di  
Jing, Xia  
Johnson, Andrew  
Johnson, Kevin  
Johnson, Kipp  
Johnson, Sharon  
Johnson, Stephen  
Johnson, Steven  
Johnson, Todd  
Jones, Josette  
Jonnagaddala, Jitendra  
Joyce, Cara  
Jung, Hyunggu  
Kabel, Margo  
Kachuee, Mohammad  
Kaelber, David
Kagawa, Rina, Keller, Michael, Kim, Namkug
Kamdar, Maulik, Kelley, Marjorie, Kim, Sun
Kandaswamy, Swaminathan, Kent, Erin, Kim, Tae
Kang, Min, Khairat, Saif, Kim, Yong-Mi
Kang, Tian, Khalifa, Yassir, Kim, Youngjun
Kannampallil, Thomas, Khan, Faiza, Kimura, Eizen
Kannan, Vaishnavi, Khan, Mirza, Kimura, Karen
Kanter, Andrew, Khan, Naqi, King, Andrew
Kanthawala, Shaheen, Khanna, Raman, Kingsbury, Paul
Kapoor, Akshat, Kharrazi, Hadi, Kirchner, Les
Karami, Amir, Khetan, Vanitha, Kirkendall, Eric
Karampourniotis, Panagiotis, Kholghi, Mahnoosh, Kistler, William
Kartoun, Uri, Khoong, Elaine, Klann, Jeffrey
Kashani, Kianoush, Khurshid, Anjum, Klassen, Prescott
Kashyap, Nitu, Kilicoglu, Halil, Kleinberg, Samantha
KASHYAP, VIPUL, Killeen, James, Kocaballi, Ahmet
Kate, Rohit, Kim, Bommae, Kochar, Preeti
Kato-Lin, Yi-Chin, Kim, Bora, Kolotylo-Kulkarni, Malgorzata
Kauchak, David, Kim, Bryan, Komorowski, Matthieu
Kaufman, David, Kim, Dokyoon, Kondylakis, Haridimos
Kayaalp, Mehmet, Kim, George, Kong, Guilan
Kebapci, Ayda, Kim, Min, Kong, Ha
Keber, Barbara, Kim, Min-hyung, Koola, Jejo
Keenan, Gail, Kim, Minjung, Koopman, Bevan
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li, Kai</td>
<td>Liu, Lin</td>
<td>Luyckx, Kim</td>
</tr>
<tr>
<td>Li, Muqun</td>
<td>Liu, Mei</td>
<td>Lyalin, David</td>
</tr>
<tr>
<td>Li, Yan</td>
<td>Liu, Nan</td>
<td>Lyle, Jay</td>
</tr>
<tr>
<td>Li, Ying</td>
<td>Liu, Sijia</td>
<td>Lyles, Courtney</td>
</tr>
<tr>
<td>Li, Yu</td>
<td>Liu, Siru</td>
<td>Lyman, Jason</td>
</tr>
<tr>
<td>Liang, Chen</td>
<td>Liu, Wanli</td>
<td>Lytle, Kay</td>
</tr>
<tr>
<td>Liang, Gongbo</td>
<td>Liu, Xu</td>
<td>Ma, Fenglong</td>
</tr>
<tr>
<td>Liang, Ou</td>
<td>Liu, Yongtai</td>
<td>Madani, Sina</td>
</tr>
<tr>
<td>Liang, Wayne</td>
<td>Liu, Hao</td>
<td>Madera, Martin</td>
</tr>
<tr>
<td>Lim Choi Keung, Sarah</td>
<td>Liu, I-ching</td>
<td>Maestre, Juan</td>
</tr>
<tr>
<td>Lin, Jia-Ling</td>
<td>Ljubic, Branimir</td>
<td>Mahajan, Satish</td>
</tr>
<tr>
<td>Lin, King-Ip</td>
<td>Loehfelm, Thomas</td>
<td>Mahdavinejad, Mohammad</td>
</tr>
<tr>
<td>Lin, Sunny</td>
<td>Loehnhardt, Benjamin</td>
<td>Mahnke, Andrea</td>
</tr>
<tr>
<td>Lin, Wei-Chun</td>
<td>Loonsk, John</td>
<td>Mahouli, Malika</td>
</tr>
<tr>
<td>Lin, Yucong</td>
<td>Lopez Campos, Guillermo</td>
<td>Major, Vincent</td>
</tr>
<tr>
<td>Lindoerfer, Doris</td>
<td>López-García, Pablo</td>
<td>Maldonado, Ramon</td>
</tr>
<tr>
<td>Ling, Albee</td>
<td>Lorenzi, Virginia</td>
<td>Malec, Scott</td>
</tr>
<tr>
<td>Ling, David</td>
<td>Lossio-Ventura, Juan</td>
<td>Malenfant, Jessica</td>
</tr>
<tr>
<td>Ling, Yuan</td>
<td>Lourie, Eli</td>
<td>Maleyeff, John</td>
</tr>
<tr>
<td>Lingren, Todd</td>
<td>Loversidge, Jacqueline</td>
<td>Malin, Bradley</td>
</tr>
<tr>
<td>Little, David</td>
<td>Lu, Chris</td>
<td>Maloney, Christopher</td>
</tr>
<tr>
<td>Liu, Dianbo</td>
<td>Lucic, Ana</td>
<td>Maloney, Francine</td>
</tr>
<tr>
<td>Liu, Jin</td>
<td>Luo, Lingyun</td>
<td>Mammen, Manoj</td>
</tr>
<tr>
<td>Liu, Lei</td>
<td>Luo, Yuhan</td>
<td>Manders, Eric-Jan</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Mandl, Kenneth</td>
<td>McClure, Robert</td>
<td>Menking, Amanda</td>
</tr>
<tr>
<td>Manilich, Elena</td>
<td>McCormick-Ricket, Iben</td>
<td>Menon, Priya</td>
</tr>
<tr>
<td>Manion, Frank</td>
<td>McCoy, Allison</td>
<td>Mensah, Nana</td>
</tr>
<tr>
<td>Manning, John</td>
<td>McCoy, Matthew</td>
<td>Metke Jimenez, Alejandro</td>
</tr>
<tr>
<td>Manos, LaVerne</td>
<td>McCune, Jeannine</td>
<td>Michael, David</td>
</tr>
<tr>
<td>Mao, Yuqing</td>
<td>McDonald, Samuel</td>
<td>Michaels, Maria</td>
</tr>
<tr>
<td>Marceglia, Sara</td>
<td>McFarlane, Timothy</td>
<td>Mikles, Sean</td>
</tr>
<tr>
<td>Marcial, Laura</td>
<td>McGarvey, Sunanda</td>
<td>Miksch, Tim</td>
</tr>
<tr>
<td>Marshall, Robert</td>
<td>McGilchrist, Mark</td>
<td>Miller, Timothy</td>
</tr>
<tr>
<td>Marsolo, Keith</td>
<td>McGraw, Deven</td>
<td>Milojevic, Stasa</td>
</tr>
<tr>
<td>Martin, Richard</td>
<td>McInnes, Bridget</td>
<td>Min, Aehong</td>
</tr>
<tr>
<td>Martinez, William</td>
<td>McKelvey Jr., Greg</td>
<td>Miñarro-Gimenez, Jose</td>
</tr>
<tr>
<td>Martínez-Romero, Marcos</td>
<td>McKinley, Danette</td>
<td>Miotto, Riccardo</td>
</tr>
<tr>
<td>Massoudi, Barbara</td>
<td>McKnight, Lawrence</td>
<td>Mishra, Rashmi</td>
</tr>
<tr>
<td>Masud, Jakir</td>
<td>McRoy, Susan</td>
<td>Mitchell, Elliot</td>
</tr>
<tr>
<td>Mathan, Santosh</td>
<td>McVay, Megan</td>
<td>Mitchell, Jason</td>
</tr>
<tr>
<td>Matheny, Michael</td>
<td>Meadows, Ginny</td>
<td>Mitsch, Christoph</td>
</tr>
<tr>
<td>Mathias, Patrick</td>
<td>Medina, Angeli</td>
<td>Mittu, Ranjeev</td>
</tr>
<tr>
<td>Matis-Mitchell, Sherri</td>
<td>Mei, Jing</td>
<td>Moen, Anne</td>
</tr>
<tr>
<td>Mayampurath, Anoop</td>
<td>Melamed, Rachel</td>
<td>Mollah, Shamim</td>
</tr>
<tr>
<td>Maygers, Joyce</td>
<td>Melton, Brittany</td>
<td>Monsen, Karen</td>
</tr>
<tr>
<td>Mazzotti, Diego</td>
<td>Mendonca, Eneida</td>
<td>Moore, Carlton</td>
</tr>
<tr>
<td>McBride, Susan</td>
<td>Meng, Frank</td>
<td>Moore, Carrie</td>
</tr>
<tr>
<td>McCleary, Nadine</td>
<td>Mengling, Feng</td>
<td>Morey, Angela</td>
</tr>
<tr>
<td>Morgan, Stephen</td>
<td>Nahm, Eun-Shim</td>
<td>Noh, Jiho</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Mork, James</td>
<td>Nanji, Karen</td>
<td>Nolen, Kim</td>
</tr>
<tr>
<td>Morris, Michele</td>
<td>Natarajan, Annamalai</td>
<td>Norden, Andrew</td>
</tr>
<tr>
<td>Morton, Stuart</td>
<td>Naumann, Tristan</td>
<td>Nordo, Amy</td>
</tr>
<tr>
<td>Mosa, Abu</td>
<td>Nebeker, Jonathan</td>
<td>Norman, Christopher</td>
</tr>
<tr>
<td>Moscovitch, Ben</td>
<td>Nemati, Shamim</td>
<td>Nosal, Sarah</td>
</tr>
<tr>
<td>Moss, Laura</td>
<td>Newbold, Susan</td>
<td>Novak, Laurie</td>
</tr>
<tr>
<td>Mossman, Ned</td>
<td>Newman, David</td>
<td>Obara, Piotr</td>
</tr>
<tr>
<td>Motiwalla, Tasneem</td>
<td>Newman-Griffis, Denis</td>
<td>Obeidat, Rand</td>
</tr>
<tr>
<td>Mougins, Fleur</td>
<td>Ng, Kenney</td>
<td>Oberhardt, Matthew</td>
</tr>
<tr>
<td>Mower, Justin</td>
<td>Ngufor, Che</td>
<td>O'Bryan, Kevin</td>
</tr>
<tr>
<td>Mowery, Danielle</td>
<td>Nguyen, Anthony</td>
<td>Odebode, Iyanuoluwa</td>
</tr>
<tr>
<td>Mueller, Lars</td>
<td>Nguyen, Hai-Son</td>
<td>Oh, Hyunkyoung</td>
</tr>
<tr>
<td>Mukherjee, Partha</td>
<td>Nguyen, Hien</td>
<td>Oh, Wonsuk</td>
</tr>
<tr>
<td>Mundt, Barbie</td>
<td>Nguyen, Nam</td>
<td>Ohkuma, Rika</td>
</tr>
<tr>
<td>Murcko, Anita</td>
<td>Nguyen, Vickie</td>
<td>Okamoto, Kazuya</td>
</tr>
<tr>
<td>Musser, Clayton</td>
<td>Nho, Kwangsik</td>
<td>Olex, Amy</td>
</tr>
<tr>
<td>Muth, Andrew</td>
<td>Ni, Yizhao</td>
<td>Omoloja, Abiodun</td>
</tr>
<tr>
<td>Muthu, Naveen</td>
<td>Nicholls, Lance</td>
<td>Omri, Mohamed</td>
</tr>
<tr>
<td>Myerburg, Stuart</td>
<td>Nicklaus, Krista</td>
<td>Ong, Irene</td>
</tr>
<tr>
<td>Myers, Risa</td>
<td>Nieves-Robbins, Neris</td>
<td>Oniki, Thomas</td>
</tr>
<tr>
<td>Nabeel, Ismail</td>
<td>NIKIEMA, Jean</td>
<td>Orenstein, Evan</td>
</tr>
<tr>
<td>Nagarajan, Radha</td>
<td>Nimley, Kenneth</td>
<td>Orwoll, Benjamin</td>
</tr>
<tr>
<td>Nahlawi, Layan</td>
<td>Ning, Xia</td>
<td>Osborne, John</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Ostovari, Mina</td>
<td>Patel, Jay</td>
<td>Philip, John</td>
</tr>
<tr>
<td>Ouyang, En</td>
<td>Patel, Jigar</td>
<td>Phillips, Mark</td>
</tr>
<tr>
<td>Oyoyo, Udo</td>
<td>Patel, Vipul</td>
<td>Phillips, Michael</td>
</tr>
<tr>
<td>Ozaydin, Bunyamin</td>
<td>Pater, Jessica</td>
<td>Piad, Sean</td>
</tr>
<tr>
<td>Ozeran, Larry</td>
<td>Patra, Braja</td>
<td>Pichon, Adrienne</td>
</tr>
<tr>
<td>Ozkaynak, Mustafa</td>
<td>Patrick, Timothy</td>
<td>Pierce, David</td>
</tr>
<tr>
<td>Ozmen, Ozgur</td>
<td>Patterson, Olga</td>
<td>Pillai, Malvika</td>
</tr>
<tr>
<td>Pacheco, Jennifer</td>
<td>Payne, Philip</td>
<td>Pirtle, Claude</td>
</tr>
<tr>
<td>Paciotti, Brian</td>
<td>Peek, Niels</td>
<td>Piscotty Jr., Ronald</td>
</tr>
<tr>
<td>Paik, Kenneth</td>
<td>Pei, Menglu</td>
<td>Plasek, Joseph</td>
</tr>
<tr>
<td>Painter, Ian</td>
<td>Peltonen, Laura-Maria</td>
<td>Podila, Pradeep</td>
</tr>
<tr>
<td>Pal, Doyel</td>
<td>Peng, Yifan</td>
<td>Poeran, Jasvant</td>
</tr>
<tr>
<td>Palchuk, Matvey</td>
<td>Pentakalos, Odysseas</td>
<td>Polifroni, Joseph</td>
</tr>
<tr>
<td>Pan, Eric</td>
<td>Pepper, Catherine</td>
<td>Pollack, Ari</td>
</tr>
<tr>
<td>Pan, Mei-Lien</td>
<td>Perera, Bede</td>
<td>Polzin, Thomas</td>
</tr>
<tr>
<td>Panahiazar, Maryam</td>
<td>Perkins, Randa</td>
<td>Poon, Eric</td>
</tr>
<tr>
<td>Papagari Sangareddy, Sridhar</td>
<td>Perl, Yehoshua</td>
<td>Poon, Simon</td>
</tr>
<tr>
<td>Park, Albert</td>
<td>Perotte, Rimma</td>
<td>Popescu, Mihail</td>
</tr>
<tr>
<td>Park, Jung</td>
<td>Perros, Ioakeim</td>
<td>Popovic, Jennifer</td>
</tr>
<tr>
<td>Park, Sulki</td>
<td>Petersen, Carolyn</td>
<td>Possanza, Lorraine</td>
</tr>
<tr>
<td>Park, Yoonyoung</td>
<td>Peterson, Kevin</td>
<td>Povalej Brzan, Petra</td>
</tr>
<tr>
<td>Parra-Calderón, Carlos</td>
<td>Pfaff, Emily</td>
<td>Powers, James</td>
</tr>
<tr>
<td>Parry, Dave</td>
<td>Pfeffer, Michael</td>
<td>Pozhitkov, Alex</td>
</tr>
<tr>
<td>Patel, Arpit</td>
<td>Pfohl, Stephen</td>
<td>Pradhan, Meeta</td>
</tr>
<tr>
<td>Roosan, Moom</td>
<td>Saboury, Babak</td>
<td>Saxe, Glenn</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Rosati, Robert</td>
<td>Safarova, Mayya</td>
<td>Scells, Harrisen</td>
</tr>
<tr>
<td>Rose, Rebecca</td>
<td>Saha, Amit</td>
<td>Scheringer, Harm</td>
</tr>
<tr>
<td>Rosenau, Paul</td>
<td>Sahoo, Satya</td>
<td>Schilling, Lisa</td>
</tr>
<tr>
<td>Rosenbloom, Samuel</td>
<td>Sakilay, Sylvester</td>
<td>Schlegel, Daniel</td>
</tr>
<tr>
<td>Rosenthal, Alex</td>
<td>Saleh, Sameh</td>
<td>Schmit, Cason</td>
</tr>
<tr>
<td>Rossi, Lorenzo</td>
<td>Salkowitz, Susan</td>
<td>Schneider, Daniel</td>
</tr>
<tr>
<td>Rouhizadeh, Masoud</td>
<td>Salmasian, Hojat</td>
<td>Schorler, Anna</td>
</tr>
<tr>
<td>Rousseau, Justin</td>
<td>Salsabili, Mahsa</td>
<td>Schreiber, Richard</td>
</tr>
<tr>
<td>Roychoudhury, Shoumik</td>
<td>Samal, Lipika</td>
<td>Schrom, John</td>
</tr>
<tr>
<td>Ru, Boshu</td>
<td>Samarth, Anita</td>
<td>Schuler, James</td>
</tr>
<tr>
<td>Rubin, Luciana</td>
<td>Samavedam, Rajni</td>
<td>Schumacher, Michael</td>
</tr>
<tr>
<td>Rubini, Juliet</td>
<td>Sangha, Veer</td>
<td>Schweig, Kelsey</td>
</tr>
<tr>
<td>Rudin, Robert</td>
<td>Sanghvi, Trisha</td>
<td>Seale, Deborah</td>
</tr>
<tr>
<td>Rudniy, Alex</td>
<td>Santos, Ricardo</td>
<td>Sebastian, Sharon</td>
</tr>
<tr>
<td>Ruiz, Carolina</td>
<td>Sanz, Javier</td>
<td>Sejdic, Ervin</td>
</tr>
<tr>
<td>Ruiz, Victor</td>
<td>Sarantis, Demetrios</td>
<td>Selva, Luis</td>
</tr>
<tr>
<td>Rule, Adam</td>
<td>Saraya, Abraham</td>
<td>Seo, Dongwoo</td>
</tr>
<tr>
<td>Rusincovitch, Shelley</td>
<td>Sarker, Abeed</td>
<td>Seppälä, Selja</td>
</tr>
<tr>
<td>Russ, Alissa</td>
<td>Sarmiento, Raymond</td>
<td>Sethi, Isha</td>
</tr>
<tr>
<td>Russell, Seth</td>
<td>Sassoon, Isabel</td>
<td>Sezgin, Emre</td>
</tr>
<tr>
<td>Rust, Laura</td>
<td>Sathyanarayanan, Shrie</td>
<td>Shachak, Aviv</td>
</tr>
<tr>
<td>Rutledge, Geoffrey</td>
<td>Savin, Michael</td>
<td>Shah, Harshal</td>
</tr>
<tr>
<td>Ryan-Lora, Beatriz</td>
<td>Sawesi, Suhila</td>
<td>Shah, Nigam</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Shah, Nirav</td>
<td>Sholle, Evan</td>
<td>Sommer, Janine</td>
</tr>
<tr>
<td>Shankar, Prabhu</td>
<td>Shrivastava, Susmita</td>
<td>Song, Lixin</td>
</tr>
<tr>
<td>Sharko, Marianne</td>
<td>Si, Yuqi</td>
<td>song, wenyu</td>
</tr>
<tr>
<td>Sharma, Ashish</td>
<td>Silva, Henrique</td>
<td>Song, Xing</td>
</tr>
<tr>
<td>Shatkay, Hagit</td>
<td>Silverman, Robert</td>
<td>Soni, Hiral</td>
</tr>
<tr>
<td>Shaukat, Muhammad</td>
<td>Simms, Andrew</td>
<td>Soni, Sarvesh</td>
</tr>
<tr>
<td>Shea, Christopher</td>
<td>Simpao, Allan</td>
<td>Soto-Campos, Gerardo</td>
</tr>
<tr>
<td>Sheets, Lincoln</td>
<td>Simpson, Christopher</td>
<td>Sottara, Davide</td>
</tr>
<tr>
<td>Sheikhtaheri, Abbas</td>
<td>Sinha, Shyamashree</td>
<td>Sow, Daby</td>
</tr>
<tr>
<td>Shellum, Jane</td>
<td>Sissung, Tristan</td>
<td>Speier, William</td>
</tr>
<tr>
<td>Shelov, Eric</td>
<td>Sittig, Dean</td>
<td>Srinivasan, Arunkumar</td>
</tr>
<tr>
<td>Shen, Feichen</td>
<td>Skapik, Julia</td>
<td>Srivastava, Avi</td>
</tr>
<tr>
<td>Shen, Yuanyuan</td>
<td>Slee, Ann</td>
<td>St. Jacques, Paul</td>
</tr>
<tr>
<td>Sherman, Eli</td>
<td>Smedley, Nova</td>
<td>Stamenova, Vess</td>
</tr>
<tr>
<td>Shi, Jianlin</td>
<td>Smith, Jaime</td>
<td>Starr, Richard</td>
</tr>
<tr>
<td>Shi, Tian</td>
<td>Smith, Kirk</td>
<td>Steidinger, Shawn</td>
</tr>
<tr>
<td>Shi, Yunfeng</td>
<td>Smith, Wade</td>
<td>Stein, Daniel</td>
</tr>
<tr>
<td>Shibahara, Takuma</td>
<td>Smith, Catherine</td>
<td>Stephens, Kari</td>
</tr>
<tr>
<td>Shickel, Benjamin</td>
<td>Snowdon, Jane</td>
<td>Stetson, Pete</td>
</tr>
<tr>
<td>Shimpri, Neel</td>
<td>Snyder, Mark</td>
<td>Stidham, Ryan</td>
</tr>
<tr>
<td>Shin, Soo-Yong</td>
<td>Soguero-Ruiz, Cristina</td>
<td>Stockwell, Melissa</td>
</tr>
<tr>
<td>Shivade, Chaitanya</td>
<td>Sohn, Sunghwan</td>
<td>Stonbraker, Samantha</td>
</tr>
<tr>
<td>Shivers, Jennifer</td>
<td>Sokolov, Jeff</td>
<td>Stone III, Norman</td>
</tr>
<tr>
<td>Shojaee Bakhtiari, Ali</td>
<td>Solomonides, Anthony</td>
<td>Strachna, Olga</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Turpin, Rebecca</td>
<td>Voorhees, Ellen</td>
<td>Wang, Taowei</td>
</tr>
<tr>
<td>Turvey, Carolyn</td>
<td>Vrana, Scott</td>
<td>Wang, Wei</td>
</tr>
<tr>
<td>Tweedy, Carolyn</td>
<td>Vu, Thanh</td>
<td>Wang, Xiaoyan</td>
</tr>
<tr>
<td>Unertl, Kim</td>
<td>Vydiswaran, V.G.Vinod</td>
<td>Wang, Yanshan</td>
</tr>
<tr>
<td>Uzun, Ece</td>
<td>Wagholikar, Kavishwar</td>
<td>Wang, Yefeng</td>
</tr>
<tr>
<td>Vallo Hult, Helena</td>
<td>Waitman, Lemuel</td>
<td>Wang, Ying</td>
</tr>
<tr>
<td>Valtchinov, Vlad</td>
<td>Walcott, Aisha</td>
<td>Wang, Ying</td>
</tr>
<tr>
<td>Van Auken, Kimberly</td>
<td>Walden, Anita</td>
<td>Wang, Yu</td>
</tr>
<tr>
<td>Van Cain, Melissa</td>
<td>Waller, Anna</td>
<td>Wang, Yuanbo</td>
</tr>
<tr>
<td>van der Veer, Sabine</td>
<td>Walsh, Colin</td>
<td>Warner, Jeremy</td>
</tr>
<tr>
<td>Van Panhuis, Willem</td>
<td>Walton, Nephi</td>
<td>Weber, Jens</td>
</tr>
<tr>
<td>Van Way III, Charles</td>
<td>Wan, Zhiyu</td>
<td>Weber, Patrick</td>
</tr>
<tr>
<td>VanHouten, Jacob</td>
<td>Wanderer, Jonathan</td>
<td>Wegier, Pete</td>
</tr>
<tr>
<td>VanZandbergen, Christine</td>
<td>Wang, Amy</td>
<td>Wehbe, Firas</td>
</tr>
<tr>
<td>Varga, John</td>
<td>Wang, Chen</td>
<td>Wei, Chih-Hsuan</td>
</tr>
<tr>
<td>Varner, Michael</td>
<td>Wang, Dongwen</td>
<td>Wei, Duo</td>
</tr>
<tr>
<td>Vazquez, Janette</td>
<td>Wang, Fusheng</td>
<td>Wei, Wei</td>
</tr>
<tr>
<td>Vemireddy, Sravani</td>
<td>Wang, Han</td>
<td>Wei, Ying</td>
</tr>
<tr>
<td>Verhagen, Marc</td>
<td>Wang, Jingqi</td>
<td>Weiner, Mark</td>
</tr>
<tr>
<td>Verspoor, Karin</td>
<td>Wang, Liqin</td>
<td>Weir, Charlene</td>
</tr>
<tr>
<td>Vickers, Andrew</td>
<td>Wang, Liwei</td>
<td>Weiskopf, Nicole</td>
</tr>
<tr>
<td>Visweswaran, Shyam</td>
<td>Wang, Lucy</td>
<td>Weiss, Jeremy</td>
</tr>
<tr>
<td>Voils, Corrine</td>
<td>Wang, Ping</td>
<td>Weissman, Gary</td>
</tr>
<tr>
<td>Volchenbouroum, Samuel</td>
<td>Wang, Shunchao</td>
<td></td>
</tr>
</tbody>
</table>

27
Welch, Brandon
Wissel, Benjamin
Yan, Xiaowei

Wen, Hongyi
Woeltje, Keith
Yang, Chengliang

West, Vivian
Wojtusiak, Janusz
Yang, Fei

Westmorland, Jennifer
Wong, Anthony
Yang, Jianji

Westra, Bonnie
Woo, Kyungmi
Yang, Jie

Wetzel, Randall
Wood, Grant
Yang, Qing

Wheeler, Terrie
Woods, Anne
Yang, Xi

Whipple, Mark
Workman, Terri
Yang, Yushi

White, Jules
Wu, Chunlei
Yao, Lixia

Whitley, Eric
Wu, Danny
Yao, Robert

Whittenburg, Luann
Wu, Julie
Yazdavar, Amir

Wickramasinghe, Nilmini
Wu, Patrick
Ye, Cheng

Wiley, Ken
Wu, Shinya
Yen, Po-Yin

Wiley Jr, Kevin
Wu, Yonghui
Yigzaw, Kassaye

Wilkinson, Katie
Wyatt, Kirk
Yim, Wen-wai

Wilkinson, Ronald
Xia, Eryu
Yoho, Amanda

Wilkinson, Thomas
Xia, Jingbo
Yousefi, Safoora

Willett, DuWayne
Xia, Weiyi
Yu, Fei

Williams, Karmen
Xie, Sherrie
Yu, Kun-Hsing

Williams, Marc
Xing, Zhaopeng
Yu, Sean

Williams, Nick
Xu, Enliang
Yu, Shao

Wills, Jonathan
Xu, Song
Yu, Xinyu

Wilson, Marisa
xu, zhenxing
Yu, Yiqin

Winden, Tamara
Yan, Chao
Yu, Yue
NOTICE

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

Julia Adler-Milstein, PhD¹, You Chen, PhD², Michelle Hribar, PhD³, Jen Popovic, DVM, MA⁴, J. Marc Overhage, MD, PhD⁵

¹University of California San Francisco, San Francisco, CA; ²Vanderbilt University Medical Center, Nashville, TN; ³Oregon Health & Science University, Portland, OR; ⁴RTI International, Waltham, MA; ⁵Cerner Corporation, Kansas City, MO

Abstract

EHR systems generate log data, which can be productively employed to improve the efficiency, quality, and safety of healthcare. There is a growing number of studies that use EHR log data but, as with any new set of data and associated methods, there is little consistency in approach and there are substantial inefficiencies from researchers replicating foundational work. Following a highly successful AMIA panel in 2018 that introduced log data and its potential uses, we convened 3 workgroups to develop measurement and methods guidance for working with log data. Our panel will present results from each workgroup and illustrate how the products could be used to support a range of research topics. The panel will also engage in a discussant-led conversation about the opportunities, limitations, and methodological challenges of common approaches, and where additional efforts may be needed. The learning objectives for this panel include: increasing awareness of the value of EHR log data for answering various clinical informatics and clinical research questions, understanding new methodological guidance for how to use these data in a research setting, and understanding the remaining domains in which community-led work could improve the quality and efficiency of EHR log data research.

Panel Description

An EHR system generates log data by recording and time stamping events and actions occurring within the system. Traditionally utilized to detect security and compliance issues, EHR log data are increasingly used to address varied clinical informatics and clinical research questions, particularly in the domains of workflow, efficiency, and patient safety¹, ². Because these data are not entered by humans and instead are an automatic byproduct of EHR use, there are fewer challenges with data quality and integrity. However, there are also unique but pervasive challenges in working with log data in a research context.

In early 2018, a community-based effort launched the National Research Network (NRN) for EHR Log Data. The network serves as a forum to connect researchers across the country working with log data and to facilitate sharing research projects, best practices, methodologies, etc (via monthly webinars). As the network rapidly grew to more than 80 participants, it became clear that there was substantial heterogeneity in approaches to tackling similar issues and many examples emerged of different researchers and sites replicating the same foundational work. In response, the network convened 3 workgroups – each taking on a different domain in which developing common methods and measures along with methodological guidance would be valuable. The first workgroup is developing a framework to characterize frontline user task types that are measurable in log data. The second workgroup is developing definitions for prevalent log data measures, such as the start and end time of an encounter. The third workgroup is facilitating an assessment of comparability of log data across sites by developing standards to capture a single, highly-specified workflow.

During the panel, the workgroup leads will present the goals and products of their workgroups as well as give examples of their application to varied research questions. The presentation of results will be followed by a discussant-led conversation, intended to stimulate thought about the strengths and limitations of the workgroup outputs, extensions
of the workgroup efforts that could increase their value, and future needs of researchers working with EHR log data. The panelists represent different institutions (VUMC, OHSU and RTI) and have diverse background and training. Specifically:

**Julia Adler-Milstein, PhD** (moderator and organizer) will kick-off the panel by introducing the panelists and the topic. Dr. Adler-Milstein will also describe the rationale and motivation for bringing this panel together and the efforts of the broader National Research Network that she leads.

**You Chen, PhD** (panelist) will present the results from the first workgroup “Developing a framework to characterize task types”. This workgroup has 19 members coming from 14 different institutions including academic and industrial medical centers, universities and EHR vendors. The goal of the workgroup is to create a framework of task types that capture common clinical activities. Audit logs record a care provider’s activities - such as “accessing a treatment plan” (view), “signing a clinical note” (modify) or “printing an operating room report” (export) - and an EHR system’s activities - such as “a quick note was automatically sent in response to an e-consult” (system). Informed by a literature review to identify commonly reported task categories, the workgroup is collecting access actions generated by different EHR vendors (e.g., Epic and Cerner), and conducting a comparative analysis to build a catalog of task types including administrative, billing and coding, chart review, documentation (notes), orders (medications, labs), patient email, etc. For each task type, the goal is to generate a definition and a list of corresponding access actions that are commonly available in log data. We will discuss limitations and challenges to creating common definitions and associated measures (e.g., differences in granularity of log data entries) as well as provide recommendations for future work that address these challenges.

**Michelle Hribar, PhD** (panelist) will present the results from the second workgroup “Developing measures relevant to audit log work”. This workgroup is comprised of 20 members from 17 different institutions including universities, healthcare centers, government, and industry, and is addressing the lack of common approaches to measure activity timing derived from audit logs. The goal of this workgroup is to produce a catalog of measures with definitions, validations, and applications across different medical settings, EHR vendors and EHR implementations. High-priority measures include those that are frequently reported in EHR log data research, such as “provider documentation time” and “exam start time”. Work on newer measures such as “shift start” and “shift end” times will also be presented along with the methodology for defining and validating these measures across settings, institutions, and EHR vendors. Limitations and challenges for defining these measures will be discussed as well as recommendations for improving their accuracy and reproducibility.

**Jen Popovic, DVM, MA** (panelist) will present the results from the third workgroup “Developing standards to support use of audit log data to characterize specific use cases”. This workgroup is comprised of a dozen members from 11 different institutions, including universities, industry, government and healthcare centers. The goal of this workgroup is to identify and leverage existing/emerging models, such as FHIR resources, to support log data standards-development around a use-case. The selected use-case is, “Characterizing patterns of EHR use across clinical departments and/or types of staff”. The panelist will discuss the resources needed to address this use-case (e.g., Audit Events, Tasks, Practitioners), describe the metrics used with those resources (e.g., Login/Logout), present differences discovered in log data capture and meaning across systems (e.g., whether timestamps associated with Login/Logout represent the same event, such as a user-submit or an authentication), and propose the standards to address the use-case. She will describe limitations and challenges to developing standards across these resources and events, as well as future work needed to scale and extend them to additional use-cases.

**J. Marc Overhage, MD, PhD** (discussant). Dr. Overhage will leverage his substantial experience with EHR log data to facilitate a participatory discussion with the panel and audience to address the challenges and opportunities regarding uses of workgroup outputs. Questions to be considered include:

- What are the primary barriers to uptake and effective use of common approaches to task categorization, time-based measures, and use cases? How can these barriers be overcome?
• How can we be confident that findings from EHR log file analyses supporting workgroup efforts reflect actual user activities?
• How can we consider EHR end-user (e.g., clinical staff) perspectives about log data use/misuse, particularly to the extent that the data may or may not accurately reflect actual user activities?
• How important is comparability of findings from EHR log file analysis with results obtained using other methods to support credibility of workgroup efforts?
• What additional work would enable more rigorous or more efficient approaches to log data analysis across institutions?

Topic Rationale
EHR log data are a potentially powerful input for clinical informatics research. To date, usage of these data demonstrates their applicability to answering a range of important clinical informatics and clinical research questions. However, methodological inconsistency threatens both the rigor and the efficiency of the burgeoning work in this area. The AMIA community is poised to be a leader in sharing guidance that helps the entire community use EHR log files to advance this promising area. It is therefore critical to share current community-led efforts to ensure broad awareness and solicit feedback about how ongoing work can be improved and extended for greater impact. The anticipated audience for this panel will be clinical informaticists at a broad range of institutions including universities, industry (in particular, EHR vendors), government, and healthcare delivery organizations, who are building EHR log data infrastructure or using EHR audit log data to answer clinical and operational research questions.

Statement of Agreement to Participate
All panelists have approved of this submission and agreed to participate.

Panel Participant Details
Julia Adler-Milstein, PhD. Director, Center for Clinical Informatics & Improvement Research. Associate Professor, Department of Medicine, University of California San Francisco
You Chen, PhD. Assistant Professor, Department of Biomedical Informatics, Vanderbilt University Medical Center
Michelle Hribar, PhD. Assistant Professor, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University
Jen Popovic, DVM, MA. Senior Director, Center for Health Data Analytics, RTI International
Marc Overhage MD, PhD. Chief Medical Informatics Officer and VP Intelligence Strategy, Cerner Corporation

References
Quantifying Efficiencies Gained Through Shareable Clinical Decision Support Resources

Shafa Al-Showk, MPH, CHES¹, Edwin A. Lomotan, M.D.¹, Kristen Miller, DrPH, CPPS², A. Zach Hettinger, MD, MS²; Jeremy Michel, MD, MHS³,
¹Agency for Healthcare Research and Quality, Rockville, MD; ²National Center for Human Factors in Healthcare, MedStar Health, Washington, DC; ³University of Pennsylvania and the Children’s Hospital of Philadelphia, Philadelphia PA

Abstract

Clinical decision support (CDS) is the process of bringing evidence into practice by providing the right information, to the right audience, using the right channel and format, at the right time (known as the “Five Rights” of CDS). Health care systems have traditionally developed CDS in siloes, which is a sign of redundancy and inefficiency of an overall ecosystem where CDS knowledge might be shared. MedStar, through contract with AHRQ, studied potential efficiencies gained by health care systems using shared, interoperable CDS available through CDS Connect, AHRQ’s web-based platform for sharing CDS artifacts. This panel describes how researchers worked with four health care systems to quantify differences in process, time, and resources between CDS developed entirely internally versus CDS developed with shared CDS artifacts. At the end, participants will be able to translate this knowledge into a business case for health care systems. Participants will be able to access CDS Connect and review its collection of shareable, interoperable clinical decision support. Participants will be able to compare and contrast processes for developing CDS in siloes vs. developing CDS using shared CDS artifacts. Participants will be able to identify measures that quantify potential efficiencies, including time and resources, towards developing CDS.

Introduction

One way to potentially accelerate the movement of research evidence into practice is through computerized clinical decision support (CDS). The process includes translating evidence-based knowledge such as clinical practice guidelines into computable form and then implementing the guidance into an electronic health record (EHR) or other technology as CDS. The design, development, and deployment of CDS is time consuming and resource-intensive. These three major components are essential in the CDS lifecycle (Figure 1). Health care systems currently develop CDS in siloes, which has been estimated at a collective cost of $25 billion, if the process is repeated across all currently recommended care within each health care system in the U.S.¹. This cost may be lowered if the process were more systematic and replicable across settings and if the process were supported by shareable, interoperable, standards-based tools and resources. These include tools for authoring CDS, structured representations of guidelines that are machine-readable, standards-based logic specifications, and implementation guidance. Health care systems would essentially not have to “start from scratch” when faced with CDS development and implementation. CDS Connect, a web-based platform developed by the MITRE Corporation under contract with AHRQ, has provided many of these tools since its inception in 2016. Health system leaders need to know how these tools may affect their current CDS processes, and in particular, whether such tools can improve the efficiency by which CDS can be developed and implemented.
Panel Description

This didactic panel will feature the leads of CDS Connect and researchers investigating potential efficiencies gained by using shareable, interoperable CDS. Panelists will focus on sharing their experience of CDS built in an isolated compared to a sharable state while discussing preliminary findings and lessons learned.

**Shafa Al-Showk, MPH, CHES**, serves as a health scientist and program official in the Division of Health IT in the Center for Evidence and Practice Improvement at AHRQ. She will provide background for the project and how it relates to AHRQ’s overall CDS initiative. Shafa will also moderate the panel.

**Edwin Lomotan, M.D.,** serves as Chief of Clinical Informatics for the Division of Health IT in the Center for Evidence and Practice Improvement at AHRQ. He will describe the CDS Connect platform, including its purpose, history, and content of interoperable CDS tools and resources. As a discussant on the panel, Dr. Lomotan will also provide reactions to results of the Medstar study, including describing potential implications for how health care systems may share CDS in the future and the role of CDS repositories.

**Kristen Miller, DrPH, CPPS,** is Scientific Director of the National Center for Human Factors in Healthcare at MedStar Health and an Associate Professor at the Georgetown University School of Medicine. Dr. Miller is co-principal investigator on the AHRQ contract to quantify efficiencies gained through shareable, interoperable CDS such as that provided through CDS Connect. She will describe analyses and findings to define and quantify efficiencies gained by health care systems to build a business case that could be used to promote the use of shared CDS within organizations (see Figure 1). The research team composed of health information technology experts, human factors experts, and a combination of clinical and non-clinical backgrounds, compiled a list of potential metrics that could be used to quantify and evaluate efficiencies of the CDS lifecycle. These metrics were used in evaluations that included stakeholder discussions and participatory data collection for CDS team members at four healthcare systems in a cross-sectional evaluation of CDS builds with and without the use of CDS Connect. The study design allows for comparisons of processes to evaluate team makeup, reduction of redundant work, and availability of resources. Dr. Miller will share qualitative and quantitative comparisons of the processes. Preliminary findings suggest shareable resources could fill the gap, which requires healthcare systems to independently develop, deploy, and manage CDS. The use of shareable CDS resources most significantly impacts the design process, reducing time and manpower required to retrieve evidence and extract content.

**A. Zachary Hettinger, MD, MS,** serves as the Director of Cognitive Informatics and Medical Director for the Human Factors Center, providing medical and informatics expertise and direction on grants and contracts while practicing clinically and working to improve clinical informatics operations across MedStar Health. Dr. Hettinger is co-principal investigator on the AHRQ contract to quantify efficiencies gained through shareable, interoperable CDS such as that provided through CDS Connect. He will describe MedStar’s experience as an academic nonprofit healthcare system that uses Cerner, going through the CDS lifecycle (see Figure 1) of the isolated CDS build versus the shareable CDS build with specific attention paid to the role of EHR vendor sponsored CDS tools.
Jeremy Michel, MD, MHS, is an assistant professor at the University of Pennsylvania and attending physician at the Children’s Hospital of Philadelphia (CHOP). Dr. Michel will relate the experience of CHOP as a pediatric health care system that uses Epic by comparing and contrasting the development of an isolated CDS build versus a shareable CDS build. He will describe his experiences using the CDS Connect platform to disseminate shareable CDS including and benefits and barriers identified while supporting external site implementations of shareable CDS.

Relevance to Annual Symposium Theme and Importance of the Topic

Our proposed panel fits directly into the theme of the AMIA 2019 Annual Symposium, “Clinical Informatics: Findings related to the design, development, and implementation and maintenance of state-of-the-art clinical systems, including electronic health records, standards and interoperability, clinical decision support, and effects on clinical quality, safety, and patient outcomes.” Health care systems spend considerable resources developing CDS but could take advantage of others’ experiences, including re-using knowledge “artifacts” generated by translating scientific evidence into computable form. Experience is growing in the use of health IT standards that support interoperable CDS and in the use of public infrastructure that supports sharing of CDS knowledge artifacts and tools.

Discussion Questions for Audience Participation

We anticipate the panel will generate discussion in a wide range of areas including business case model and process for CDS development. The panel will also explore with the audience how CDS Connect may serve organizations and patients. Discussions questions may include:

- What are some of the challenges health care systems faced while developing CDS in isolation?
- Does the form of CDS (e.g. semi-structured vs. computer-interpretable structured representation) influence the efficiencies gained by the health care system?
- How might the artifacts and the authoring tool enable you to develop CDS more efficiently?
- How can the CDS build in the isolated state differ from the CDS build in the shareable state?
- How would your health care system approach the option of using CDS artifacts developed elsewhere but available for sharing?

Author statement

All authors have agreed to take part in the panel.

References

Translating, Implementing, Deploying, and Evaluating Clinical Interventions Using Machine Learning Based Predictive Models: Illustrative Case Studies

Yin Aphinyanaphongs, MD/PhD₁, Jonathan Wilt², Corey Chivers, PhD³, Mark Sendak, MD⁴

₁NYU Langone, New York, NY; ²Oschner Health System, Jefferson Parish, LA
³Penn Medicine, Philadelphia, PA, ⁴Duke Institute for Health Innovation, Durham, NC,

Abstract

Significant barriers to the translation of machine learning based predictive models in clinical care continue to exist. The barriers are both technical and social and successful deployments require significant effort. This panel will provide targeted case studies of machine learning based predictive models through model evaluation, deployment, and adoption. The panel will demonstrate end to end examples of machine learning based predictive models and the issues and processes leading to their successful deployment. In some case studies, model evaluation results demonstrate improvements in quality of care and reductions in cost of care.

Learning Objectives

After participating in this session, attendants should be able to:

- Demonstrate examples of machine learning based predictive models deployed and adopted in clinical practice.
- Describe challenges in successful deployment and adoption of machine learning based predictive models.
- Describe operational principles supporting the deployment of machine learning based predictive models.
- Describe methods for evaluating the clinical impact of these machine learning based predictive models.
- Describe interventions and measurement tools at the point of care.
- Describe how organizational and engineering structure contribute to model translation.

General Description

Precision medicine (1) and precision delivery (2) are two emerging approaches to healthcare delivery. Both efforts propose to target care delivery to those who can most benefit from an intervention. At its core, this targeting task is what every healthcare provider does when a patient walks into our offices. We gather patient data and allocate our resources (medicines, diagnostics, etc) to uncover and address underlying pathology. Precision medicine and precision delivery propose to expand this paradigm with omics and clinical data respectively.

Over many decades, researchers published papers and models with excellent discriminatory performance on multiple clinical problems. In the clinical domain, these models have typically been described as calculators. For example, PORT (3) and Reynolds (4) are two clinical calculators. Some models have been widely adopted such as SIRS (5). However, far more that are published do not. Multiple barriers may account for the limited adoption. One particular barrier was the difficulty in entering data into these calculators. Simple calculators that do not require many predictors are simple to use though they may lack discriminatory power due to their simplicity. Calculators that may require significant numbers of predictors are non-starters for practical clinical use. With the advent of electronic health records and digital recording devices, this barrier is lower and it is now possible to build larger models and potentially more performant models. However, the quality of data stored in the EHR and difficulty extracting and normalizing data to a common data model introduces new challenges to effectively scaling machine learning applications across settings. In addition, the cost to validate and integrate an analytics application is not trivial (6). Concurrently with the digitization of health data, gains in pattern recognition and machine learning have enabled significant improvements in model performance. We are once again at a crossroads with many models being published and few being deployed.

Our premise is that barriers for pragmatic model implementation still exist between the academic research that demonstrates excellent discriminatory performance for multiple clinical problems and the translation of these machine learning based predictive models into routine practice. Naively it would seem easy to believe that
Translating a published model into practice is as easy as deploying a model on a server and feeding it variables and reporting the outcome. In practice, the challenges to deployment and adoption of models are not trivial.

This panel will highlight the lifecycle of a machine learning based predictive model through the experiences of panelists implementing, deploying, and evaluating models at their respective institution. Through case studies, we will each highlight issues and challenges associated with deployment and adoption of machine learning based predictive models.

**Description**

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 minutes</td>
<td>Aphinyanaphongs</td>
<td>Predictive analytics in medicine and the predictive model pipeline.</td>
</tr>
<tr>
<td>12 minutes</td>
<td>Aphinyanaphongs</td>
<td>Implementation, deployment, and integration of a 2 month mortality model in clinical practice.</td>
</tr>
<tr>
<td>15 minutes</td>
<td>Chivers</td>
<td>Application of a 6 month mortality risk model to direct inpatient palliative care resources.</td>
</tr>
<tr>
<td>15 minutes</td>
<td>Wilt</td>
<td>Integrating a real-time patient deterioration predictive model directly into clinical workflows.</td>
</tr>
<tr>
<td>15 minutes</td>
<td>Sendak</td>
<td>Translating sepsis machine learning model into practice.</td>
</tr>
<tr>
<td>25 minutes</td>
<td>Aphinyanaphongs</td>
<td>Panel Discussion with Audience.</td>
</tr>
</tbody>
</table>

Each panelist will anchor their content with a case study of a predictive model that they have built, deployed, or evaluated at their institution.

**Aphinyanaphongs:** Dr. Aphinyanaphongs is the Director of Clinical Predictive Analytics at NYU Langone Medical Center. His job is to build, deploy, monitor, and evaluate machine learning based models for optimizing clinical operations and care. His team includes data scientists and software engineers that build sustainable infrastructure to translate clinical needs and research findings into hospital solutions. Dr. Aphinyanaphongs will moderate the session and discuss the overall lifecycle of a predictive model. He will discuss NYU’s experience in deploying a 2 month mortality prediction model at the point of care.

**Chivers:** Dr. Chivers is a Senior Data Scientist at Penn Medicine where he works with clinicians to take predictive healthcare solutions from the idea and experimentation phases all the way to scaled, production implementations. His background is in computational biology and applied statistical modeling with a focus on the management of complex systems and decision making under uncertainty. Dr. Chivers will discuss a current Penn Medicine project using a 6 month mortality risk model to direct inpatient palliative care resources with specific attention to issues surrounding the observability/ measurability of joint prediction-intervention systems in a health care setting.

**Wilt:** Mr. Wilt is the Chief Technology Officer of innovation at Ochsner, an innovation lab and accelerator founded by Ochsner Health System in 2015 to reimagine and revolutionize the experience and delivery of healthcare. He leads the product development and is responsible for the overall technology strategy, advanced analytics and AI, and data integration. He will discuss the successes and challenges of implementing a real-time inpatient deterioration predictive model directly into clinical workflows in Epic.

**Sendak:** Dr. Sendak is the Population Health & Data Science Lead at the Duke Institute for Health Innovation, where he leads interdisciplinary teams of data scientists, clinicians, and machine learning experts to build technologies that solve real clinical problems within DukeHealth. He leads a medical student innovation scholarship and data science internships to train the workforce required to bring medicine into a new digital age. He will discuss
the successful implementation of an inpatient sepsis prediction model.

Significance of panel topic and intended audience

Data science and predictive analytics are taking increasingly large roles in health care. A plethora of research has demonstrated that machine learning based predictive models can identify patients for care prioritization and potentially reduce costs (7). Translating these findings into actual practice is essential to realize their potential gains. Healthcare organizations have recognized this potential and are making investments in building infrastructure and teams. Thus, this panel topic is timely and needed. Our intended audience include machine learning researchers who want to translate their papers into practice, data scientists within healthcare organizations, healthcare executives who want to assess infrastructure and conceptual concerns in making investments in predictive analytics, change managers and behavioral economists who can address the challenges of model adoption, user interface designers who build the end users experiences, medical educators who train future health care providers, and clinicians who will design and implement the interventions based on the model recommendations.

Discussion Questions

- What technical challenges did you overcome?
- How did you communicate with the clinical team to drive change?
- What tools and systems support successful predictive model translation?
- What resourcing (in people, software, and hardware) is necessary for successful translation?
- What skillsets are necessary for translational clinical predictive analytics teams.
- What mechanisms are possible to pool our collective knowledge and experiences in deployments to identify best practices?

Participation Statement

All proposed panelists have agreed to participate in the panel if the proposal is accepted.

References

Abstract

This panel will introduce the Informatics Technology for Cancer Research (ITCR) program, an NCI initiative to support investigator-initiated informatics technology driven by critical needs in cancer research. Investigators supported by ITCR will share their perspective on the importance and challenges in developing and disseminating informatics technology stemming from an academic setting, using their own projects as case studies, and will discuss how programs like ITCR can benefit these efforts. Panelists will also describe how their tools are benefiting the scientific goals of the targeted cancer research communities and how interactions with end users have shaped the development of their tool. In addition, the panel will discuss ways in which the cancer informatics community of practice that is evolving through ITCR is promoting collaboration and interoperability within and beyond program participants. The NCI ITCR program coordinator will moderate the discussion and will also provide a brief overview of NCI’s goals for supporting informatics technology development.

A general description of the panel and issues that will be examined

Remarkable progress in biomedical informatics supporting cancer research has been made in recent years. The emerging use of cloud computing to support big data analysis; innovative computational methods for variant calling and driver mutation detection; the application of machine learning and other artificial intelligence methods to image analysis; and the application of natural language processing to clinical information, for example, have all had fundamental impacts on cancer research. However, the use of informatics in everyday research remains a challenge for a number of reasons including the low availability of user-focused tools; lack of sufficient support for existing tools; the need for training materials; lack of interoperable tools; and the lack of tools to meet specific scientific needs. These critical needs are recognized by NCI and formed the basis of the NCI Informatics Technology for Cancer Research (ITCR) Program whose scope serves informatics needs that span the cancer research continuum.

In this panel, four ITCR investigators will describe their informatics development projects and how involvement in the ITCR community has impacted their efforts. These projects represent a range of cancer informatics activities, including cancer genome visualization, network-based treatment decision support, medical image analysis, and electronic health record (EHR) data exploration. Each investigator will outline the scientific goals and impact of their technology and discuss their perspective on informatics technology development in an academic setting, including how programs like ITCR can support their efforts. Topics for consideration will include user outreach, engagement, and support; resource sustainability; interoperability; security; and scalability. The ITCR program coordinator will moderate the panel and will also provide perspective on the NCI’s investments in informatics technology development and future goals for this program.

An explanation why the topic of this panel is timely

Cancer research is a data intensive science and accelerating progress in this field demands information technology to support the collection, management, integration, and analysis of cancer research data, especially as cancer research is entering a “big data” era. For technology to truly support and impact science, it must be driven and guided by research priorities. In addition, successful tools will also be user friendly, stable, secure, and interoperable with other tools in the research workflow. This panel is aimed at addressing the important goals and challenges with developing informatics technology that can have the broadest impact on cancer research.
Panelist’s perspective

Simina Boca, Ph.D. will discuss the CDGnet tool, which provides therapy recommendations for precision oncology using biological networks. For many cancer types, it is becoming increasingly common to check for specific molecular features in order to decide on a targeted treatment plan. Our objective is to expand the range of targeted therapy options for cancer patients who undergo molecular profiling by developing a user-friendly, evidence-based approach that accounts for the cross-talk within cancer pathways and is personalized for the individual patient. CDGnet fulfills this goal by incorporating a variety of publicly-available resources and combining them into biological networks.

David Hanauer, M.D. will discuss EMERSE (project-emersere.org), which is a free text search engine for clinical concepts in the EHR. EMERSE has been used for a wide variety of cancer research tasks ranging from cohort identification to data abstraction. ITCR support has been instrumental to build the infrastructure necessary for widespread dissemination, allowing the team to focus on security, documentation, training, and outreach. Ongoing ITCR-supported efforts also include developing new functionality for efficiently querying clinical notes, and securely networking the instances for cross-institutional cohort discovery.

Ian Holmes, Ph.D. will discuss JBrowse (jbrowse.org), a fast JavaScript/HTML5 genome browser that can be deployed as a static site with very low infrastructural overhead. JBrowse has been used to visualize genomic variants in cancer, for example by the UK Cancer Genome Project project. With ITCR support, the project is building cancer informatics features into JBrowse, including a circular viewer for analyzing the distribution of structural variants, a richly integrated human dataset, and modern web tooling to facilitate the construction of bioinformatics web dashboards.

Despina Kontos, Ph.D. will discuss the cancer imaging phenomics toolkit (CaPTk), a dynamically growing software platform for analysis of radiographic images of cancer. CaPTk integrates advanced tools leveraging machine learning and medical image analysis algorithms that have been developed in the context of active cancer imaging studies to derive phenotypic imaging signatures of diagnostic and predictive value. With emphasis given on being a very lightweight and efficient image viewer and eliminating the prerequisite for a substantial computational background, CaPTk thus aims to facilitate the swift translation of advanced computational algorithms into routine clinical quantification, analysis, decision making, and reporting workflow. Results from clinical studies in brain, breast, and lung cancer will be demonstrated.

Juli Klemm, Ph.D. will provide a brief overview of the program and NCI’s goals for funding investigator-initiated informatics technology development. She will moderate the panel and focus discussions on challenges and best practices for successful informatics technology development for cancer research.

Discussion questions to enhance audience participation

1. How do you engage the user community to ensure your tool is usable by and useful to the target audience?
2. What are the challenges for taking software from a research project to something that is hardened, generalizable, and usable across a wide variety of settings?
3. What are the common challenges that software projects encounter that ITCR funding helps to overcome?
4. What roadblocks to tool adoption and dissemination have you experienced?
5. What are key features for a software to succeed in an impactful clinical translation?
6. What are the barriers to incorporating emerging software tools in cancer clinical trials?

Agreement to participate

All speakers have agreed to attend AMIA 2019 Annual Symposium and participate on this panel.


Toward Patient-facing Clinical Decision Support: Critical Issues and Near-term Opportunities

Aziz Boxwala MD, PhD; Janet Desroche, Blackford Middleton, MD, MPH, MSc; Joshua E. Richardson PhD, MS, MLIS; Julie A. Scherer, PhD
1Elimu Informatics, Inc. Richmond, CA; 2MEDITECH, Westwood, MA; 3Apervita, Inc. Chicago, IL; 4RTI International Chicago, IL; 5Motive Medical Intelligence, San Francisco, CA

Abstract
A confluence of factors is changing patient-facing clinical decision support (CDS) from a possibility to a reality. Recent product advances, the growing adoption of wearable consumer technologies, and policy updates are but some of the many factors to open CDS beyond hospital and clinic EHRs to consumers wherever they go. How is patient-facing CDS to best support patients and how will any resulting data from patients inform clinicians’ decision-making or support learning health systems? This panel is comprised of both pioneers in the field as well as current implementers and will consider critical issues as to how the marketplace is addressing CDS for patients in light of rapidly changing interoperability standards, rapid shifts in the healthcare and consumer markets, and game-changing policies such as the 21st Century Cures Act and the Office of the National Coordinator’s Notice of Proposed Rulemaking. Through field experiences and clinical scenarios, the audience will come to understand the challenges and opportunities for patient-facing CDS and learn new approaches to developing and implementing such as the AHRQ-funded Patient-Centered CDS Learning Network’s efforts to pilot a patient-facing CDS demonstration application and candidate reference architecture.

Significance
Clinical decision support (CDS) has historically been the realm of clinicians in healthcare settings where knowledge is delivered at the “right” opportunities to alert, remind, or deliver current evidence about a patient and/or population. Yet importantly, CDS is rapidly moving from the domain of electronic health records (EHRs) into tools (e.g. smartphones, home computers, voice assistants, etc.) that interact directly with patients and even their caregivers. This shift from EHR-based CDS to potentially ubiquitous CDS presents game-changing implications on ways that informaticians, policy-makers, and vendors address the challenges (and opportunities) that “patient-facing CDS” poses.

The Patient-Centered CDS Learning Network defines patient-facing CDS as, “CDS that involves direct participation by individual patients in their health-related decisions and actions by leveraging reliable and timely evidence from patient-centered outcomes research findings and/or patient-specific information (e.g. patient-generated health data).” There are multiple technological and standards-related challenges for effectively integrating patient-facing CDS into care pathways. Recent policy announcements and programs highlight the timeliness – if not urgency – of the topic. The National Institutes of Health’s 21st Century Cures Act, the Office of the National Coordinator’s Notice of Proposed Rulemaking (NPRM), and the Food and Drug Administration’s Software as a Medical Device (SaMD) are just some of many recent policy efforts that have implications for patient-facing CDS. Furthermore, technology is rapidly advancing to the point that wearable trackers and devices are attaining research- and medical-grade levels of validity and reliability. It will soon be feasible – maybe even common - for patient-directed CDS tools to be made available. How are clinicians to incorporate these data into workflows and use these data to inform their own decision-making and communications with patients?

It is key for audiences to understand the ways that current (and rapidly evolving) policies, standards, and technologies will soon be changing the ways that next generation products, services, and research will need to incorporate patient-
facing CDS. This panel includes presenters with multiple perspectives – including developer, EHR vendor, content provider, and research – as to how current the next generation of tools and services will address these systematic changes.

Description of the Panel

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5’</td>
<td>Richardson</td>
<td>Moderator – Introduction of the panelists</td>
</tr>
<tr>
<td>15’</td>
<td>Boxwala</td>
<td>Review of patient-facing CDS and considerations related to interoperability including the HL7 SMART-on-FHIR architecture</td>
</tr>
<tr>
<td>15’</td>
<td>Scherer</td>
<td>Considerations (including security and privacy) for delivering CDS content to consumers</td>
</tr>
<tr>
<td>15’</td>
<td>Desroche</td>
<td>Considerations for integrating patient-generated health data and patient-facing CDS into EHRs</td>
</tr>
<tr>
<td>15’</td>
<td>Middleton</td>
<td>Considerations of policies and technologies for maintaining trust in patient-facing CDS</td>
</tr>
<tr>
<td>25’</td>
<td>All</td>
<td>Discussion, Q&amp;A with audience</td>
</tr>
</tbody>
</table>

Learning Objectives

1. Understand the major barriers, and opportunities, to wider use of patient-facing CDS in practice
2. Learn about new approaches to implementing and delivering patient-facing CDS
3. Learn about standards for CDS and progress of the Learning Network Patient-facing CDS Work Group

Individual Speaker Contributions

**Joshua Richardson, PhD, MS, MLIS** – Dr. Richardson will serve as the panel moderator and introduce the panel participants as well as frame the discussion. Later in the program, Dr. Richardson will summarize the takeaways from the panelists and lead the audience Q&A.

**Aziz Boxwala, MD, PhD** – Dr. Boxwala will briefly review the history of personal health records and clinical decision support and then describe how these areas are merging into one area: patient-facing clinical decision support. He will provide insights into how the HL7 SMART-on-FHIR and other CDS standards can enable patient-facing CDS.

**Julie Scherer, PhD** – Patient-facing CDS at scale requires longitudinal patient data, computable clinical knowledge, and sophisticated knowledge execution services. Dr. Scherer will discuss design, development, and implementation requirements for patient-facing CDS; explore the issues associated with achieving scale while ensuring privacy; and examine available patient-facing CDS solutions.

**Janet Desroche** – Patient-facing CDS poses multiple challenges and opportunities to EHR platforms such as ways any clinician may “prescribe” CDS for patients via smartphones, and conversely, incorporate the patient-reported outcomes that result. As an EHR vendor executive, Ms. Desroche will discuss how EHR designs and implementations will increasingly account for patient-facing CDS and solutions for supporting shared decision-making between patients and providers. She will draw examples from her experience as a member of the College of Healthcare Information Management Executives (CHIME) Opioid Task Force.

**Blackford Middleton, MD, MPH, MSc** – Dr. Middleton will briefly review recent policy and technology developments driving patient-facing CDS and describe work from the Patient-Centered CDS Learning Network on maintaining Trust in decision support artifacts. He will offer suggestions for how these principles translate into practice for various forms of knowledge repositories.
Expected Discussion and Discussion Questions
Interest in patient-facing CDS and in implementing provider-facing CDS continues to grow. However, the adoption of provider-facing CDS is increasing slowly at best and may be eclipsed by the adoption of consumer-facing CDS at scale in the near term. We expect the audience will want to engage the panelists in discussions on approaches for effective implementation of patient-facing CDS, approaches to standards-based approaches to patient-facing CDS systems. The discussion questions include:

1. How are informatics standards evolving to meet the requirements of patient-facing CDS in research and healthcare more broadly?
2. In what ways are content providers addressing patient needs for reliable and actionable CDS?
3. In what ways are EHR vendors adapting to support patient-facing CDS and consumption of patient-generated health data?
4. How will current CDS approaches (people, processes, and technology) need to evolve to promote adoption within the culture of medical care?

Intended Audience
- **CMIOs and CNIOs:** responsible for implementing clinical decision support in healthcare organizations, and considering patient-facing CDS
- **CMO/CQO/CNO:** responsible for achieving high quality care, reduced unwarranted clinical variation, reduced patient harm, patient engagement strategies
- **EHR Implementers:** implementation staff responsible for implementation and effective use of EHR tools, and system optimization, including patient-facing tools
- **CDS Systems Developers:** development staff building clinical decision support tools and services for providers and their patients
- **Knowledge vendors** supplying content for CDS services
- **Policy authors:** those setting healthcare policies and creating clinical practice guidelines

Attestation
The panel moderator has assurances from all participants that they will be available to participate at AMIA 2019.

References
If You Build It, They Will Come:

The National Patient-Centered Clinical Research Network (PCORnet):
From Conception to Execution

Tom Carton, PhD, MS1, Maryan Zirkle, MD MS MA2, Shayna Barbash2, Betsy Shenkman, PhD, MSN3, Abel Kho, MD4, Adrian Hernandez, MD, MHS5
1Louisiana Public Health Institute, New Orleans, LA; 2Patient Centered Outcomes Research Institute, Washington, DC; 3University of Florida, Gainesville, FL; 4Northwestern Feinberg School of Medicine, Chicago, IL.; 5Duke Clinical Research Institute, Durham, NC

Abstract
The Patient-Centered Outcomes Research Institute (PCORI) funded the National Patient-Centered Clinical Research Network, PCORnet. PCORnet leverages the power of electronic health data in a national distributed research network (DRN) to conduct patient-centered research. From late 2013 to present, PCORnet has built, refined and expanded its infrastructure to connect various stakeholders in productive research partnerships that leverage the power of large volumes of high-quality health data maintained by partner networks. This panel will focus on the unique attributes and lessons learned in the areas of governance, research capabilities, and sustainability during PCORnet’s evolution toward a cohesive self-governing Network.

This panel will describe PCORnet’s advancement across governance, research infrastructure and overall network sustainability.

Learning Objectives:
At the conclusion of this panel, participants will be able:
1. To describe the PCORnet infrastructure and how it has changed over time.
2. To discuss the lessons learned conducting research in PCORnet.
3. To describe the process for accessing PCORnet resources and the long-term business model for leveraging and sustaining PCORnet’s innovative infrastructure.

Panel breakdown

Moderator: Maryan Zirkle

The goal of this panel is to present information on PCORnet’s evolution toward a sustainable Network that conducts high-quality Patient-Centered Research (PCR). The panel will include four panelists each describing unique attributes and lessons learned from the Network’s infrastructure development and implementation.

I. PCORnet Overview (Panelist: Tom Carton)

The panelist will provide information on the building blocks of PCORnet. The presentation will serve as a foundation for understanding the Network’s distributed infrastructure and operations.

II. PCORnet Governance and Partnerships (Panelist: Betsy Shenkman)

Over the years, PCORnet has evolved into a cohesive, self-governing Network that connects various stakeholders through efficient governance. The presentation will highlight PCORnet’s unique decision-making processes and describe how various partners work in collaboration to support sustainability.

III. PCORnet Research Capabilities (Panelist: Abel Kho)
PCORnet has been utilized in dozens of research studies. The presentation will provide information on significant results from specific research studies utilizing PCORnet and infrastructure refinements in the following domains: data quality, operational techniques/processes, and stakeholder engagement.

IV. PCORnet Sustainability: Front Door and Business Development (Panelist: Adrian Hernandez)

The final portion of the panel will focus on PCORnet’s shift to a new business model in support of sustainability from various funders and organizations. The presentation will describe the process for accessing PCORnet resources as well as the overall business development process.

Discussion Questions:

1. What are some of the components of a clinical research network that have proven to be integral in a network’s successful and long-term sustainability?
2. What business models have demonstrated to be successful in national networks?
3. Are there additional research capabilities that should be considered within PCORnet?
4. What areas of governance should be further strengthened within the Network?
5. What additional partnerships should be considered within the Network?

Impact and Targeted Audience:

The current clinical research ecosystem faces several challenges in conducting research, including engagement of key stakeholders and generation of timely evidence that answers questions that matter most to patients and their doctors. PCORnet helps to overcome these challenges through using an infrastructure, tools and policies that support efficient clinical research, utilizing multiple data sources, and enabling people and systems to work collaboratively. This panel provides an overview of the opportunities to collaborate and engage with PCORnet partners, and showcases a unique governance structure that underlies a Network that has and will continue to make a significant impact on the health care ecosystem.

This panel will be targeted at clinicians, researchers, educators, healthcare leaders, policy makers, and Health IT professionals who have had little or no contact with PCORnet.

There are no prerequisites for this panel, but a basic understanding of data modeling or distributed research networks would be helpful.

Statement on panel participation:

All panelists included in this panel have agreed to participate on the panel.
ONC’s Leading Edge Acceleration Projects (LEAP) Advance an Application Programming Interface-based Ecosystem with Real World Use Cases for Population Health and Patient Care

Kevin Chaney, MGS\(^1\), Teresa Zayas-Cabán, PhD\(^1\), Kenneth D. Mandl, MD, MPH\(^2\), Kristen Miller, DrPH, CPPS\(^3\)

\(^1\)Office of the National Coordinator for Health Information Technology, Washington, DC; \(^2\)Computational Health Informatics Program, Boston Children’s Hospital, Boston MA; \(^3\)National Center for Human Factors in Healthcare, MedStar Health, Washington, DC

Abstract

It is critical the field of health IT innovate far quicker to optimize real-time solutions in areas ripe for momentum. The 21st Century Cures Act requires health IT to have application programming interfaces (APIs) which enable access to all of a patient’s data with no special effort. The Office of the National Coordinator for Health Information Technology (ONC) published a proposed rule to implement these APIs, both for bulk data access and for one at a time patient access. In anticipation of a new generation of health IT, ONC created the Leading Edge Acceleration Projects (LEAP) in Health IT funding opportunity, informing the development, implementation and use, and/or advancement of well designed, interoperable, and scalable health IT. Two projects under LEAP in health IT, will be showcased on this panel. Dr. Mandl will discuss and demonstrate a provider-payor use case employing the emerging SMART/HL-7 bulk-data API, which defines a Fast Healthcare Interoperability Resources (FHIR®) “Flat FHIR” format. Dr. Miller will present on work advancing clinical knowledge at the point of care via risk calculator implemented as a SMART on FHIR app into an active surveillance tool clinicians can use in electronic health records (EHRs) to integrate and improve patient care.

Introduction

The Office of the National Coordinator for Health Information Technology (ONC) drives the administration’s efforts to increase the adoption of health IT and promote nationwide interoperability of health information to improve health care\(^1\). In 2010, through the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009\(^2\), ONC created the Strategic Health Information Technology Advanced Research Projects (SHARP) cooperative agreement program\(^3\) to close the gap between the promise of health IT and its realized benefits. Substitutable Medical Applications, Reusable Technologies (SMART) Health IT\(^4\), a standards-based technology platform that enables innovators to create apps that seamlessly and securely run across the healthcare system, is hailed as a major achievement from the SHARP program, gaining national prominence\(^5-6\).

Since then, the health care ecosystem and technology supporting it have rapidly progressed. More providers than ever have implemented electronic health record (EHR) systems and sophisticated health IT tools and applications have come to market. As the electronic exchange of health information (EHI) has matured, the amount and types of health data available has expanded too. Data standards such as Health Level 7 International’s (HL7®) Fast Healthcare Interoperability Resources (FHIR®) and application programming interfaces (APIs) are making it easier for the sharing of health data. Yet, there is still much to learn about delivering and presenting emergent data to providers and between clinical and non-clinical settings in support of care and research – especially in ways that are unobtrusive and seamless. While working to implement the 21st Century Cures Act\(^7\) provisions, ONC identified gaps with respect to leveraging EHR data to support population-level analyses and services as well as integrating clinical knowledge into routine clinical practice. The reasons for these gaps range from lack of data standards and interoperability to digitization, integration, and presentation of new evidence into clinical workflows in safe, useful, and useable ways. Without strong anticipatory research and development, these challenges could compound as we move into an era of exponential data growth.

Leading Edge Acceleration Projects (LEAP) in Health IT
In 2018, ONC published the Leading Edge Acceleration Projects (LEAP) in Health IT funding opportunity to further a new generation of health IT development and inform the innovative implementation and refinement of standards, methods, and techniques for overcoming major barriers and challenges in the field. This panel presents in detail the focus of LEAP and the ongoing work of the first two funded projects. Key aspects of both projects focus on reduction provider and health system burden of utilizing health IT, while incorporating data access and use via an API for innovative purposes. Each panelist will leave time for discussion around the development, implementation, and use of their respective solution, as well as early legal and policy implications. This panel aims to stimulate a highly interactive discussion and strengthen the community’s knowledge of innovative uses of health IT, barriers and solutions for use, and areas ripe for future work.

Panel Objectives and Presenters

The aim of this panel is to provide an overview of how ONC’s LEAP in Health IT program is preparing the U.S. for an interoperable, modular, health IT ecosystem. Panelists will discuss current and emerging challenges facing the field, demonstrate prototypes of canonical use cases, and outline priorities for next steps.

Mr. Kevin Chaney (moderator and organizer), is a Senior Program Manager at ONC and co-leads ONC’s the LEAP in Health IT program. He will introduce the session, provide an overview of ONC’s relevant portfolio of work, and describe the goals of the LEAP in Health IT program. He will also moderate the session.

Dr. Teresa Zayas-Cabán (presenter), is the Chief Scientist at ONC and her division leads ONC’s scientific and evaluation efforts including LEAP in Health IT, patient-centered outcomes research, and precision medicine programs. She will provide an update of ONC’s priorities and activities, how those align with LEAP in Health IT priorities, and emerging challenges for further exploration.

Dr. Kenneth D Mandl (presenter), directs the Computational Health Informatics Program at Boston Children’s where he leads the transformative SMART Health IT initiative and is Principal Investigator of an ONC LEAP in Health IT project. He will provide an overview of a population health use case for the FLAT FHIR Bulk API, which can drive change for populations, payers, providers, and patients at scale. Mandl is developing and testing at production scale, an open source, reference population health app for use between payers and hospitals. The project relies on the emerging SMART/HL7 bulk data export standard, which has already been implemented in EHRs through the Argonaut process, and is currently being balloted through HL7. The Centers for Medicare and Medicaid, Google and Microsoft are already using the standard. For measurement of value and performance, providers and payors can share a well-defined view into a live shared dataset in “flat FHIR” format.

Dr. Kristen Miller (presenter), is the Scientific Director of the National Center for Human Factors in Healthcare at MedStar Health, an Associate Professor of Emergency Medicine at Georgetown University School of Medicine, and Associate Faculty at the Innovation Center for Biomedical Informatics at Georgetown Medical Center. She is the Principal Investigator of an ONC LEAP in Health IT project. She will provide an overview of MedStar Health’s work transforming a stand-alone Million Hearts Risk Calculator into an interactive surveillance tool, as a SMART on FHIR app, and will discuss lessons learned in the technical design and integration, legal and policy implications, and opportunities for future enhancements. Clinical decision support is often limited by its proprietary nature specific to vendors and it is rarely integrated into clinical workflows, largely limited to a “pop-up” functionality that is largely ignored by the clinician. The innovations of the Mobilizing Million Hearts tool reimagine decision support to advance knowledge at the point of care, optimize care coordination, engage patients as partners in their care, and provide feedback regarding value-based care and risk-based accountable care organization success metrics.

Panel Discussion Questions

- Does ONC see any current or emerging areas ripe for future LEAP in Health IT?
- Are there disruptive technologies the field should be prepared for?
- What technical and policy needs or gaps, have been identified through your work limiting the usage of these FHIR-based APIs?
- Are there anticipated health IT infrastructure and/or standards barriers impeding the utilization of these solutions?
- How easily can these solutions be applied to other similar aspects in the field (eg, other risk calculators, or other bulk data types)?
Panel Learning Objectives

- Participants will have a strong understanding of ONC’s scientific efforts, with a strong focus on the LEAP in Health IT Program, and the newest focal areas of interest.
- Participants will learn about the technical development and utilization of each LEAP project to date.
- Participants will learn what challenges and barriers each LEAP project experienced and how those could impact broader uptake by the field.
- Participants will learn what gaps currently exist providing bulk data sharing and integration into real time decision making.

Conclusion

There is a growing anticipation that advances in interoperability and innovative health IT solutions will transform health care in a myriad of ways. However, as with the use of any innovative technology, emerging technical challenges and policy barriers could hobble progress. Understanding these impediments and their root causes, while grappling these issues head on through innovative approaches, LEAP can demonstrate creative ways to leverage and modernize health IT to improve health care. This panel aims to stimulate a rich discussion and gather participant input that will inform and strengthen the innovative work being explored.

Statement of Participation

Each of the panelists and the moderator have confirmed that they will participate if this submission is accepted, at the assigned timeslot during the Annual Symposium.

References

Challenges of deploying Computable Biomedical Knowledge in real-world applications
Derek Corrigan, PhD\textsuperscript{1}, Vasa Curcin, PhD\textsuperscript{2}, Jean-Francois Ethier, MD CM PhD\textsuperscript{3}, Allen Flynn, PhD\textsuperscript{4}, Davide Sottara, PhD\textsuperscript{5}
\textsuperscript{1} Royal College of Surgeons in Ireland, Dublin, Ireland. \textsuperscript{2} King’s College London, London, United Kingdom. \textsuperscript{3} Université de Sherbrooke, Quebec, Canada. \textsuperscript{4} University of Michigan, Ann Arbor, MI, USA. \textsuperscript{5} Arizona State University, Phoenix, AZ, USA

Abstract
The popularity of health data science and persistent problems with avoidable medical errors combine to put a bright spotlight on the growing need for as much actionable health knowledge as possible to be expressed in standardised, computable forms that can be shared, reused, validated, traced as it evolves across time (provenance) and extended by other researchers. Recognizing this need, several major computable biomedical knowledge organizing initiatives are underway around the globe - Research Objects, MCBK, TRANSFoRm, CodeOcean, Knowledge Grid. This panel shall deliver on four key learning objectives. Attendees will come to: (1) Understand major (a) research, (b) clinical, (c) population health, and (d) educational use cases for actionable Computable Biomedical Knowledge (CBK); (2) Recognize common barriers to the widespread adoption and use of CBK to improve health; including the need to position CBK artifacts in relation to existing standards, such as FHIR, BPMN, DMN, PMML, PFA, RDF/OWL, etc. and be able to discuss strategies for achieving large-scale CBK artifact implementation in concrete scenarios, including the routine use of CBK created outside an organisation with possibly different technologies; (3) Understand a series of key design considerations that come into play when creating CBKs. (4) Learn the challenges of managing large numbers of routinely evolving CBK artifacts in ways that span research, clinical, health operations, and educational environments and be aware of emerging software tooling used to address CBK management at scale; (5) With the idea that all knowledge is relational by nature\textsuperscript{[1,2]} in mind, envision the most critical relationships amongst CBK artifacts and how these relationships can be used to improve the findability, accessibility, interoperability, and reusability (FAIRness) of CBK artifacts, collections of CBK artifacts, and related CBK-enabled IT services.

The panel shall look into the key challenges of deploying CBK artifacts in the real world context, organised into four sessions:

Session 1: Challenges Designing Highly-functional CBKs (Flynn)

Making CBKs Findable, Accessible, Interoperable, Reusable (FAIR). From the perspective of good library practices, emerging solutions for making CBK artifacts findable, accessible, interoperable, and re-usable will be presented and remaining gaps exposed for attendees to learn from\textsuperscript{[3]}

Principles for Determining Scope of CBKs. The presentation will share a set of principles for assessing and deciding on an appropriate scope for individual CBKs and for collections of CBKs. These principles will address state and statelessness, pure functions vs. functions with side effects, and metadata to indicate the scope of CBKs and how it was determined.

Mutability and Versioning of CBKs. A series of design decisions related to managing change for CBKs will be surfaced and discussed. The topics of semantic versioning, versioning strategies, and publishing immutable snapshots of CBKs will be covered. The costs of managing CBKs and collections of CBK will be highlighted in this context.

Multiple Representations and Using CBKs in Combination. Some CBKs are better expressed using a combination of programming and/or knowledge representation languages. This poses a challenge at the integration level when the CBKs are deployed. Examples of combining multiple CBKs for effect will be shown.

Session 2: Managing CBKs (Sottara/Curcin)

Lifecycle of a CBK. The lifecycle of CBK artifacts used for Clinical Decision Support and other purposes needs to be better understood. Discrete stages of development and deployment of Clinical Prediction Rules - derivation/validation/impact analysis/dissemination - each have their associated methodologies, actors, outputs and metrics \textsuperscript{[4]}, which can be generalised to apply to CBKs as well.
Management of CBKs at scale. Deploying large quantities of CBKs (ultra-complex models; models of models; large or massive collections) is a challenge. Two components are present here: 1) the "management" at rest: how to track what all the CBKs are, how they relate and where/when are they used; 2) the "in motion" component: how do multiple CBKs deployed in the same environment interact: are all and only the ones that matter deployed, do they work in parallel/empower/conflict with each other? If they need to interact, should it be by orchestration or choreography?

Persistence and replay. To support ongoing learning in learning health systems, there are scientific and legal needs to persist CBK and to be able to "re-run" or recreate scenarios using CBK on demand. This raises issues of digital preservation, provenance/pedigree, versioning, logging/auditing, access, and technical support, among other things.

Session 3: Using CBKs (Corrigan / Ethier)

Tooling. Guideline Producing Organizations (GPOs), such as UK’s NICE, and other professional organizations are used to producing human-readable forms of content such as spreadsheets, PDFs, journal articles, white papers, etc. New tools are needed by these organizations if they are to transition successfully to producing and publishing CBK artifacts.

Interoperability. One challenge is to recognize that interoperability is a matter of degrees. Two things are interoperable according to the degree to which they function jointly to achieve a goal. Significant unresolved issues pertain to CBK interoperability, including data-level integration (for standardized, structured I/O) but also deployment-level integration (can I execute the CBK in an artifact with the tools that I have; containerization, etc.) and protocol integration (can I formulate requests and responses in a standard way - FHIR, REST-ful web services, SOAP etc.)

Session 4: Evaluating CBKs (Flynn / Ethier)

Safety and Effectiveness: As the complexity of computable biomedical knowledge (CBK) increases beyond the simple rules and formulas we use today, what must we do to ensure that it is safe and effective? How should CBK be tested? How should biases in CBK be denoted and communicated? What data about CBK must we routinely collect to study its safety and effectiveness in practice? What usage guidelines and alert mechanisms are needed to safely use CBK in practice to shape health decisions? How to measure the impact of CBK / What scenarios are more amenable to the introduction of CBK?

Applicability: In what context is this CBK intended (not) to be used? What is the purpose of a CBK module? Is it safe/possible, and if so how, to port a CBK from an environment in which it has been proven to a new one? Can/Should the information produced by a CBK be used for a different purpose?

Discussion questions

Following the panelists’ presentation of the topics above, we will invite the audience to speak about:

- Their experiences of CBKs in practice: phenotype portals, analytic script libraries, vendor repositories
- Strategies for validation of CBKs across research groups
- Lessons learned so far in unpublished work, so that the community can learn from one another.
- Reflections on the issues raised in the presentations

Panelists

The panel members are:

Derek Corrigan, PhD is an IT professional who moved into academic research with 15 years of industry experience as both a software developer and management consultant. He joined the HRB Centre for Primary Care Research working on a collaborative European Union funded research project developing infrastructure to support a Learning Health System (www.transformproject.eu). His research primary research interests include knowledge representation using ontologies and semantic web technologies, the development of computable representations of clinical evidence, service based decision support systems for primary care. Other research interests include development of OpenEHR archetypes and the implications for research of software as a medical device.
Vasa Curcin PhD (moderator) is a Senior Lecturer (Associate Professor) in Health Informatics at King’s College London. A computer scientist by training, his research focuses on reproducible informatics methods for analytics, and clinical trials and decision support embedded in Electronic Health Records, with special interest in Learning Health Systems. He is leading the Biomedical Informatics Group at the School of Population Health and Environmental Sciences.

Jean-Francois Ethier, MD CM, PhD is a clinician-scientist and associate professor in the Department of Medicine of the Université de Sherbrooke and the Sherbrooke University Health Center. His research focuses on methods to support learning health systems (LHS) to support integrated research, knowledge transfer and care delivery. More specifically, his projects target semantic interoperability between health and research infrastructures, temporal data operations to support complex health data queries as well as formal models (ontologies).

Allen Flynn PhD. Following his graduation as a Doctor of Pharmacy from the University of Michigan, Allen studied Computer Science. After working as a network analyst, he returned to pharmacy practice as a hospital pharmacist. He then became involved with several major health IT projects and developed expertise in Electronic Health Records and medication system safety. This real-world experience inspired Allen to dedicate the rest of his professional career to improving the knowledge-delivery capabilities of health IT. He went back to school and earned his doctorate at the University of Michigan School of Information. Presently, Allen is an Assistant Professor and he studies knowledge infrastructure for health while serving as the technology lead for the Knowledge Grid team of the Department of Learning Health Sciences at the University of Michigan Medical School.

Davide Sottara PhD is Principal Knowledge Engineer at Mayo Clinic and is affiliated with Arizona State University. His work focuses on research, design and development of standards-compliant, knowledge-based and model-driven clinical architectures and applications.

All of the panelists have agreed to take part in this panel and have contributed to its conceptualization and the development of this abstract.

Envisaged audience:

It is expected that the audience will be made up both of individuals who understand the need for computable representations of biomedical knowledge, and those who are new to the field - researchers, informaticians, and guideline designers.

References

Complexities of a patient-powered data era: BYO (bring your own) genomics and wearables

Erin L. Crowgey, PhD\textsuperscript{1}, Matthew K. Breitenstein, PhD\textsuperscript{2}, Theresa Walunas, PhD\textsuperscript{3}, and Samuel L. Volchenboum, MD, PhD\textsuperscript{4}

\textsuperscript{1}Nemours Alfred I. duPont Hospital for Children, Wilmington, DE; \textsuperscript{2}University of Pennsylvania, Philadelphia, PA; \textsuperscript{3}Northwestern University, Chicago, IL; \textsuperscript{4}University of Chicago, Chicago, IL

Abstract
Genomic testing has lead to a paradigm shift in our understanding and treatment of disease, with particularly noteworthy accomplishments in oncology. Recent advances in consumer genomics services such as 23andMe have empowered the patient to sequence their genome independent of their healthcare provider. In parallel, personal wearables [e.g. Garmin, FitBit] have become inexpensive and ubiquitous, providing information for personal health maintenance and activity monitoring. While patients are increasingly empowered with new information about their health, genome, and healthcare risks, there are also growing concerns about how people will access these services, how they will interpret and use the data, and how data will be protected from nefarious uses. As these products and services expand, it is critical that informaticians be amongst the most informed in the healthcare ecosystem. The learning objectives for this discussion are: 1) understand risks and benefits to patients and providers for the different approaches for generating personalized genomics data, 2) learn about strategies for engaging patients and consumers in research and return of results in a patient-driven genomic era, and 3) learn data governance, privacy, and data standards required for these large-scale projects.

Introduction
Medicine is rapidly expanding to include new and exciting ways to measure and monitor health and disease. Increasingly, information about a patient’s health is being collected outside of the traditional patient-provider interactions. New and exciting ways are emerging for patients to monitor their health and wellness and to predict the onset of disease. Traditionally, patients would have to approach their healthcare provider to obtain genomic testing, usually performed in a CLIA-certified lab. Over the past few years, the emergence of relatively-inexpensive and easily-accessible personalized genomics services has empowered people to seek and obtain their own genomic testing, outside of the traditional confines of a clinic or hospital interaction. As the costs have fallen and availability of these tests has improved, the technology continues to mature, giving patients ever-increasing access to information about their genomic information and consequent predictions about health and wellness.

In addition to personalized genomics services like 23andMe, consumers are able to collect myriad of other data related to their health. Consumer wearables and sensors provide further insights for patients about their activity and fitness. Patients often see this new wealth of information as a positive influence on their health. Many have used these tools to make positive changes to their lifestyle, for instance, modifying behavior as a result of a genomic report. Yet, the widespread availability of these data sources present challenges to both patients and providers. It is now common for patients to bring to their providers data that they have collected, either through a wearable or sensor or from a personalized genomics service. Sometimes, patients are concerned about findings and want to get their provider’s input on their risk assessment. Too often, providers are ill-equipped to field patient concerns, because these data are often poorly documented or described or the relevance to the patient’s current status is unclear. There are many opportunities now to help patients leverage these data. Under the right circumstances, this information can be a powerful way to empower patients to be invested in their own health. This panel will explore several important areas related to this new patient-driven genomics era.
Instructors

Dr. Matthew K. Breitenstein, PhD (panel moderator), is an Instructor of Informatics at the University of Pennsylvania, Perelman School of Medicine. Dr. Breitenstein is an expert in translational bioinformatics and deploys ‘flipped science’ in a full-spectrum of translational research projects (from clinic-based populations to molecular hypotheses). Dr. Breitenstein is recognized as an expert in pharmacogenomics and was recently awarded the prestigious 2019 Marco Ramoni Distinguished Paper Award for Translational Bioinformatics by the American Medical Informatics Association. The Breitenstein Lab’s methodologic research sets the stage for multi-omic data to inform ‘personalized’ or ‘N of 1’ therapeutic prescribing recommendations in hematologic and oncologic care settings.

Dr. Erin L. Crowgey, PhD, is the Associate Director of the Biomedical Research Informatics Center at Nemours Alfred I. duPont Hospital for Children. Dr. Crowgey has expertise in molecular biology and bioinformatics and is the chair of AMIA’s Genomic and Translational Bioinformatics working group. Dr. Crowgey’s primary research focus is developing and applying computational pipelines for analyzing clinical genomic data from rare Mendelian diseases and pediatric cancers. She has analyzed over >1000s of NGS libraries across multiple different assays. She will present on the history of genomic technologies, with a focus on next generation sequencing, and applications in the medical field.

Dr. Theresa Walunas, PhD, is an Assistant Professor of Medicine and Associate Director of the Center for Health Information Partnerships at Northwestern University. Dr. Walunas is an immunologist and medical informaticist who seeks to bridge clinical data, genomics and real-world data to improve outcomes using patient centered strategies. She will discuss the intersection of genomic technologies with clinical and other forms of real world data to impact outcomes and strategies for involving consumers and patients in project design and the return of results.

Samuel L. Volchenboum, PhD, MD, is an Associate Professor of Pediatrics and the Director of the Center for Research Informatics (CRI) at the University of Chicago. Dr. Volchenboum will discuss patient-driven data collection, including wearables and sensors and personalized genomics services. He will discuss the data governance, privacy, and security considerations for how these data are handled by companies and consumers. He will also discuss methods by which consumers can best leverage these data to promote their own health and wellness.

Relevance

There are different approaches for generating large scale genomic resources. The panelists will explore the similarities and differences between commercial entities such as 23AndMe, and NIH funded models including All of Us, eMERGE, and TOPmed, and foundation medicine. These different models present different risks and benefits to the consumer, and the complexities of high-throughput genomics data adds yet another level of dimensionality. This topic is relevant for patients, providers, and clinical researchers.

The panelists will also discuss the evolution of how genomics data are used for both research and business perspective purposes. In reality no data are “free,” and consumers are either paying for a service or having their data leveraged or commercialized. Interpreting genomics data is heavily dependent on analyzing an n=1 in context with pedigrees and population frequencies and demands a large-scale publicly available resource. Currently, there are several publicly-available resources, such as dbSNP, dbGap, and ClinVar, that provide a wealth of data and have enabled the identification of numerous disease associated mutations. Understanding the integration across all of these projects will be extremely valuable to the clinical research community.

Collectively, the panelists will review real-world experiences within the topics outlined. The panel consists of a multidisciplinary team with diverse expertise. Dr. Volchenboum, MD, PhD, will be able to interact with audience from a physician point of view, and also as leader in data standards and governance. Dr. Walunas intersects with the bench research and translational research worlds and will reflect on the intersection of data across these spaces as well as strategies for engaging consumers in research. Dr. Crowgey has extensive background in molecular biology, bioinformatics and genomics, and will provide expertise to the discussion in regards to data analysis and difficulties with genomic data interpretation and application, and complexities associated with returning results outside of a traditional patient-doctor interaction.
Discussion Questions

1. How are social determinants influencing this new era? Certainly, not everyone can pay for fitbits and 23AndMe? How easy should it be for patients to have their external data used as part of their routine healthcare?

2. Is the patient-centered data era a public science experiment? Where are the places in the research endeavor that patients and consumers can and should play a key role?

3. How do insurance companies play a role in the new patient-driven genomic era? How should the legal system protect patients’ privacy?

4. How does it influence a consumer when they receive genomic results without the support of a trained medical staff team? There are numerous positive examples, but it is important to also recognize the risks.

5. What are the impacts of predictive versus prognostic data, and variants of unknown significance? The differences in the application of these markers complicates the interpretation and application. Are patients and providers prepared?

Declaration

Submitting author, Erin L. Crowgey, confirms that the panelists agree to attend the Symposium in November if this abstract is accepted for presentation.

Acknowledgements

ELC would like to thank the National Institute of General Medical Sciences of the National Institutes of Health grant number U54-GM104941 and the Nemours Biomedical Research Department for support.
Modernizing Health IT in Indian Country: Promoting a Patient-Centric Community Health Model

Theresa A. Cullen, MD, MS1,2, Maia Z. Laing, MBA3, Carolyn Crowder (Inupiat)4, Paul G. Biondich, MD, MS1,2, Howard Hays, MD, MSPH5

1Global Health Informatics, Regenstrief Institute, Indianapolis, IN; 2Indiana University School of Medicine, 3Office of the Chief Technology Officer, Immediate Office of the Secretary, U.S. Department of Health and Human Services, Washington, DC; Indianapolis, IN; 4Crown Consulting & Management Services, Anchorage, AK; 5Hays Health Informatics Consulting, LLC, Tucson, AZ

Abstract

The Indian Health Service (IHS) has enjoyed a 50-year history of successful, innovative and open source development of health information technology (IT). Due to multiple factors, the IHS is currently engaged in evaluating options for health IT system modernization. The Office of the Chief Technology Officer (OCTO) at the Department of Health and Human Services (HHS) collaborated with IHS and the Office of the National Coordinator (ONC) to design and conduct a multi-faceted, expert-driven and research-based approach to evaluate alternatives for health IT modernization. This panel will describe the drivers for IHS health IT modernization, the design and execution of the HHS/IHS Health IT Modernization Research Project, and methods used to ensure community engagement in modernization efforts. The panel will focus on sharing lessons learned and rubrics developed for HIT modernization applicable to other healthcare organizations.

Introduction and Background

The IHS is the principal healthcare provider for American Indian and Alaska Native (AI/AN) people in 37 states. The agency is unique among federal government healthcare systems, providing cradle-to-grave care to a defined population — members of 573 federally recognized tribes — not as a statutory mandate but as a treaty obligation. AI/AN people inhabit some of the most isolated, rural, and spectacularly beautiful parts of the United States. This population has some of the highest disease burdens and the lowest life expectancy of any ethnic group in America. The IHS has a mission encompassing personal, population and public health, and is required to report on the status of this mission to the Administration and to Congress1,2,3.

As health care and health care delivery has changed, healthcare organizations have faced the dilemma of whether, when and how to modernize the systems supporting their operations4. The breadth of Indian Health Service mission, combined with the geopolitical diversity of Indian country, creates unique challenges for health IT design and delivery1,5,6,7. These unique challenges are faced by other resource-constrained health care systems who are focused on helping their members achieve health equity by leveraging HIT solutions. This presentation discusses a comprehensive approach to modernization decision making: the development of a comprehensive qualitative, quantitative and research-based platform to identify and evaluate approaches to modernization, with a reliance on human centered design processes. The evaluation of HIT modernization for the Indian Health Service included researchers, cutting-edge technology innovators, veterans of healthcare in Indian country, and icons of health IT leadership in the U.S.- brought together to help answer how to approach health IT modernization in the IHS. Frameworks developed, lessons learned, and alternative analyses applicable to multiple organizations will be presented during this session.
Panel Description

Executing the Project – Theresa Cullen -- Presenter

Dr. Cullen will discuss the IHS Health IT Modernization Research Project in detail, describing the multiple concurrent work streams (site visits, data calls, literature research and evaluation, Human Centered Design methodology, legacy system assessment, Community of Practice modeling, and analysis of alternatives), the expertise brought to the project, and the collaboration among the project teams and stakeholders, including government agencies, tribal representatives, and community members. She will review the outcomes and recommendations of the project, including reusable frameworks developed, lessons learned and opportunities for further research. She will then introduce the panelists. At the conclusion of the panelists’ presentations, Dr. Cullen will summarize the findings of the project and how HHS and IHS are expected to proceed with subsequent phases of a multi-year modernization effort.

The HHS Perspective of the Project – Maia Laing

Ms. Laing will provide background on the IHS Health IT Modernization Research Project from the perspective of HHS, in particular that of the Office of the Chief Technology Officer. She will discuss how the project was conceived in collaboration with IHS and ONC leadership, and describe the expectations and deliverables.

History of Health IT in the IHS – Howard Hays

Dr. Hays will briefly review the 50-year history of health IT in the IHS, describing its origins on the Tohono O’odham Reservation in southern Arizona and its evolution to a comprehensive suite of over 100 applications supporting the full range of clinical and business processes at facilities serving AI/AN people across the country. He will discuss the current technical and enterprise architecture of health IT deployment in Indian country, outlining the many factors that are driving the need for modernization. This presentation will articulate the timeliness and criticality of a thoughtful evaluation of IHS options for health IT modernization\(^8,9,10\).

Tribes, Tribal Self-Determination, and Health IT – Carolyn Crowder

Ms. Crowder will orient the audience to the uniqueness and complexities of the government-to-government relationship between Indian tribes and the U.S. government. She will describe the prerogatives of Indian self-determination and how the exercise of those prerogatives by the tribes impacts programs, functions, services and activities offered to Indian people. Finally, she will offer a uniquely personal perspective as both a tribal member and tribal healthcare executive on the impact of health IT in Indian country and the importance of tribal engagement in the health IT modernization process.

Regenstrief Institute’s Take on the Problem – Paul Biondich

Dr. Biondich will discuss how Regenstrief Institute’s expertise in global health informatics was leveraged to design the approach to the IHS Health IT Modernization Research Project. He will discuss the establishment and utilization of a Technical Advisory Commission (TAC) of leading health IT academicians and executives, and the approach to evaluate and assess modernization options.

Why is this Important?

The IHS and Tribal health delivery systems have always focused on community and public health with a unique focus on holistic wellness and patient-directed clinical decision-making. IHS health IT evolved to support the same. Health IT modernization at the IHS has been driven by the increasing complexity and costs of maintaining legacy systems, an ever-changing regulatory environment, and the impact of plans by the VA to transition away from support for certain components on which IHS health IT is dependent\(^11\).

The effort to modernize the IHS health IT system seeks to harness emerging technologies to design robust, adaptable solutions that support evolving models of care. These care platforms must be designed to support patient and community engagement in a continuing effort to reverse decades of health and health care disparity faced by AI/AN people. These circumstances have provided an opportunity to evaluate IT system design and modernization options for the largest rural health network in the US. The panelists will present the approach and the rubrics developed to support the current HIT modernization evaluation. Attendees with interests in underserved and under-resourced populations, as well as those focused on modernization and future-proofing of health IT, will find this to be a compelling conversation.
Discussion Questions
1. What can be learned from the current and envisioned HIT model in Indian country which might influence future health IT development decisions for other health systems?
2. How might emerging technologies over the next decade impact the recommendations about HIT modernization?
3. How were Human Centered Design principles integrated into this project?
4. What scientific method was used to ensure that community engagement and site visits reflected an accurate perspective of the issues, gaps and requirements of users and facilities in the field?
5. How might national health policies redefine health IT system solutions to effectively address population health in the future?

Panel Organizer Statement: All participants have agreed to take part in the panel and discuss the topics as outlined above.

References
Using Social Media to Study Mental Health Conditions - Challenges and Opportunities

Vasa Curcin, PhD¹, Jyotishman Pathak, PhD², Elizabeth Ford, PhD³, Goran Nenadic, PhD⁴,⁵

¹ School of Population Health and Environmental Sciences, King’s College London, London, United Kingdom. ² Weill Cornell Medicine, Cornell University, New York, USA. ³ Brighton and Sussex Medical School, Brighton, United Kingdom. ⁴ University of Manchester, Manchester, United Kingdom. ⁵ The Alan Turing Institute, London, United Kingdom

Abstract

Social networks are increasingly seen as a promising source of data for mental health research. The resulting applications are manifold, from improving diagnostics and monitoring patient activity, to developing digital interventions and targeted advertising of mental health services. This panel will deliver on some key learning objectives in the area: (1) Understand the ethics and governance frameworks; (2) Understand recommended technology for collecting data from popular social networks (Twitter, Facebook, Reddit); (3) Discuss approaches to complex concept definitions needed for working in the mental health domain; and (4) Learn about current state-of-the-art NLP and machine learning techniques for processing social network data, their capabilities, and what the future may hold for this field.

Introduction

According to World Health Organization (WHO), the total number of people with depression globally was estimated to exceed 300 million in 2015. It is further estimated that approximately 16% of residents in the United Kingdom experience depression at some point in their lives. Likewise, in the US, approximately 16 million adults, equaling 6.7% of all adults in the country, have experienced a major depressive episode in the past year. Mental health problems are also predicted to cost 16.3 trillion USD between 2011 and 2030, through services, treatments, and a decline in productivity at work.

Collecting reliable data on mental health disorders and treatment outcomes, however, is a challenging task. For their disorder to be correctly ascertained, patients need to recall how they felt and what happened to them in the previous time period, which helps clinicians obtain comprehensive background information. However, this method is time-consuming and error-prone as sometimes patients may not be able to correctly recall their experiences. Additionally, voluntary recruitment into research studies may systematically bias the results, with those who are experiencing the most severe symptoms least likely to participate in research.

With massive volumes of data being generated on social media platforms, researchers are increasingly studying how these data relate to user behaviour, their mood and symptoms related to mental health condition. Studies have found that by focusing on user-generated messages, it is possible to screen for users with depression. This alternative method of detecting mental illness from produced content can supplement the traditional diagnostics based on recollection data, and can offer deeper insights into users past activities, experiences, behaviours, and feelings.

The panel will share recent work, challenges and opportunities that lie ahead in this area of investigation by focusing the discussions across four complementary themes:

Ethics, governance and people’s attitudes (Ford)

The key ethical issue to be determined when designing a research study using social media, is whether social media users are “participants” in the research, or whether their content is public data which stands alone. Purely observational research (no participant contact) of publicly available data (no log-in or permission needed to view), which is used for research for public benefit, and does not involve vulnerable groups, could theoretically be conducted without ethics approval or individual consent. However, it is important not to lose public trust in the endeavours of research, given
than in an age of scandals such as Cambridge Analytica, public distrust is running high. We will present the general ethical consensus around use of social media data along with guidelines for new studies.

**Data collection from social media (Curcin)**

Designing instruments to perform data capture from social networks, such as Twitter, Facebook and Reddit, is a non-trivial task. In addition to getting relevant ethics approvals where necessary and developing tools within the constraints of each platform’s API, care has to be given to the internal privacy and security rules, and these have to be followed through the lifecycle of the study to ensure that changes in regulations do not adversely affect the data collection instrument. We shall illustrate this point with a Facebook app designed to capture mental health data through CES questionnaires presented to users, and through a Twitter harvesting tool which uses hashtags to identify relevant posters and extracts their full post histories.

**Concept definitions (Nenadic / Ford)**

Identifying mental health conditions and outcomes can be challenging from user generated text. An exemplar case study will be presented to initiate discussion; this focuses on defining patient non-adherence to depression medications, reported in patient discussion forums. While named entities such as anti-depressants are straightforward to identify, determining the indication for the medication (the underlying disorder) may be much harder, as it is less frequently explicitly mentioned. Non-adherence to medication is a complex concept for which the definition needs the logical combination of several pieces of information. These issues may result in very complex annotation schema, poor inter-annotator agreement, and consequently limited accuracy in NLP algorithms.

**Techniques to analyze social media data (Pathak)**

Analysing information about mental health from social media offers some of the most challenging problems to the developers of natural language processing (NLP) and machine learning (ML) algorithms. Some evidence suggests ML algorithms may outperform rule-based NLP algorithms in this rapidly evolving environment, but they are less transparent and may exhibit hidden bias. Furthermore, in addition to textual data, many social media sites offer users the ability to post multimedia content, e.g. emojis, pictures, videos and URL links. All of these sources may communicate relevant information about the user’s state of mind. Building models from multimodal data is an important research challenge to address in improving our ability to use social media data for mental health research.

**Discussion Topics:**

Following four presentations detailing the discussion themes above, and using a range of exemplar case studies which illustrate the key issues, we will invite the audience to participate in discussions on:

- Their experiences in using social networks to collect research data
- The impact of changing governance regulations on study progress and governance approaches in different jurisdictions
- How NLP and machine learning techniques for social media data will evolve and which avenues look most promising
- Anecdotal and systematic evidence on attitudes to data mining social networks for health research
- Lessons learned so far in unpublished work, so that the community can learn from one another.

**Panelists:**

The panel members are:

- Vasa Curcin PhD (moderator) is a Senior Lecturer in Health Informatics at King’s College London. His research focuses on reproducible informatics methods for analytics, clinical trials and decision support embedded in Electronic Health Records, with special interest in Learning Health Systems.
- Elizabeth Ford PhD is a Senior Lecturer in Primary Care Research at Brighton and Sussex Medical School in Brighton, UK. Her research focuses on analytical methods, ethical issues and public engagement for
using NHS patient data and social media data for research, particularly focusing on mental health and
dementia.

- Goran Nenadic, PhD, is Professor of Computer Science at the University of Manchester and Fellow of the
Alan Turing Institute. His research focuses on natural language processing for healthcare applications. He
leads the UK healthcare text analytics network (Healtex).

- Jyotishman Pathak, PhD is the Frances and John L. Loeb Professor of Medical Informatics and Chief of the
Division of Health Informatics, Department of Healthcare Policy & Research at Weill Cornell Medicine.
His research focuses on developing and applying informatics methods for data mining and phenotype
extraction from electronic health records, and their applications in pharmacogenomics, comparative
effectiveness research, and population health research, particularly focusing on major depressive disorders.

All of the panelists have agreed to take part in this panel and have contributed to its conceptualization and the
development of this abstract.

**Intended Audience**

This panel session will appeal to all researchers interested in, or already using, social media data for mental health
research. Using a range of interesting exemplar research cases, it will introduce novice participants to key ethical
and logistical concepts for designing and executing social media studies. It will also allow more experienced
researchers to share current challenges and to discuss future directions for the research field as a whole.

**References**

March 11, 2019.
2016.
doi:10.1145/3110025.3123028
The IZ Gateway: Reducing Burden While Improving IIS Completeness Nationwide

Names of Presenter(s)/Facilitators

- Presenter 1:
  - Name: Jim Daniel
  - Academic Degree: MPH
  - Affiliation: Office of the Chief Technology Officer
  - Location: Washington, DC
  - Email: james.daniel@hhs.gov

- Presenter 2:
  - Name: David Lyalin
  - Academic Degree: Ph.D.
  - Affiliation: CDC/DDID/NCIRD
  - Location: Atlanta, GA
  - Email: dil8@cdc.gov

- Presenter 3:
  - Name: Charlie Ishikawa
  - Academic Degree: MSPH
  - Affiliation: Kahuina Consulting
  - Location: Boston, MA
  - Email: charlie.ishikawa@kahuina.com

- Presenter 4:
  - Name: Denise Chrysler
  - Academic Degree: JD
  - Affiliation: Network for Public Health Law, Mid-States Region
  - Location: Ann Arbor, MI
  - Email: dchrysler@networkforphl.org

Abstract

The goal of IZ Gateway project is to simplify the process for a multistate provider organization to submit immunization information to the appropriate immunization information system (IIS) without requiring the provider organization to create individual point-to-point connections with each IIS in the regions they operate in. This will reduce the cost for multistate provider organizations to send immunizations to the appropriate IISs and will improve the completeness of information in IIS.

The project team conducted an environmental scan to inform the development of clear technical requirements necessary to support simplified provider reporting to immunization registries across jurisdictional boundaries (without the need to make multiple connections). Organizations such as large multistate public and private provider organizations, hospital groups, adult focused care groups, large and independent pharmacy chains, and other entities involved in sending and receiving immunization data across jurisdictional boundaries were interviewed to inform this project.

Panel Description

Healthcare organizations providing immunizations to patients in multiple states are challenged to participate in modernized immunization information system (IIS) exchange, because the burden of
establishing and maintaining connections with the many IIS jurisdictions where they immunize patients is too great. Reducing this barrier and increasing multi-state provider reporting will tremendously benefit patient care, everyday public health and the effectiveness of interventions during emergencies and disasters.

The Immunization (IZ) Gateway is a Centers for Disease Control and Prevention and Health and Human Services Office of the Chief Technology Officer project to improve IIS information completeness by reducing the cost of immunization data exchange with multiple IISs for multi-state provider organizations; e.g., regional health care systems, retail pharmacies, and health service units of the Veterans Administration and Department of Defense. The project began in late-2018 with an environmental scan. Covering topics ranging from technical barriers and solutions to legal and policy challenges, the scan found issues that concern centralized onboarding, interstate immunization information sharing, and architectural and business requirements for a nationwide IZ Gateway service.

During this panel, participants will have an opportunity to learn about the IZ Gateway use cases and environmental scan findings, and discuss solutions for key technical and policy barriers. By the end of the workshop, participants will be able to...

1. describe the benefits of an Immunization Gateway service to providers, patients and public health,
2. define the legal, policy and regulatory issues involved in developing a centralized hub to share data between multi-jurisdictional providers and IIS or among IIS,
3. describe efforts that have informed the Immunization Gateway approach
4. constructively critique immunization gateway service requirements based on present IIS capabilities,
5. devise potential solutions to address legal, policy and regulatory barriers, and
6. identify an action that they can take within 3 months to advance the gateway initiative
7. leverage findings from the Immunization Gateway to apply lessons learned to other health data exchange initiatives.

**Discussion Questions:**

1. What value do stakeholders perceive in the services that an Immunization Gateway (IZ Gateway) could provide?
2. What technical solutions or architectures might ease policy barriers to multijurisdictional exchange of IIS data?
3. How do data security concerns and requirements impact the feasibility of an IZ Gateway service?
4. For an IZ Gateway-like service to be useful on a nationwide, what should be an acceptable level of tolerance for idiosyncratic IIS business practices (e.g., on-boarding and data quality monitoring)?
5. What existing information services, systems, or standards could be leveraged to accelerate multijurisdictional exchange of IIS data?
6. What other use cases might an IZ Gateway service be extended?
7. What are the scalability issues inherent to the IZ Gateway, and how might they be addressed?
8. What do you envision as the first milestone marker or indicator of success in demonstrating the IZ Gateway concept?
9. What additional benefits may the centralized hub bring beyond those discussed for immunization information systems that may be leveraged for different data exchange opportunities?
Citizen Science: Using Informatics to Engage Vulnerable Populations in Scientific Research

George Demiris PhD, FACMI1; Anne Turner MD, FACMI2; Sarah Iribarren RN, PhD2; Katherine Sward PhD, RN3

1University of Pennsylvania, Philadelphia, PA; 2University of Washington, Seattle, WA; 3University of Utah, Salt Lake City, UT

Abstract

Citizen science refers to the collection and analysis of data by various population groups in real-world settings as part of a collaboration with scientists. Citizen science aims to respect and facilitate contributions from community members to scientific inquiry, build capacity, facilitate community-driven questions and promote dissemination. The advancement of wearable and passive monitoring technologies and direct-to-consumer genetic test kits and the continued diffusion of social media have led to an increase in citizen science projects in health care. Informatics tools introduce opportunities for large cohort research studies to understand disease and drug response and the role of environmental parameters. Vulnerable populations, namely communities with limited access to resources or at a higher risk to experience discrimination, are less likely to engage in citizen science. Informatics tools for communication, data collection and engagement can address some of the challenges that vulnerable populations experience in engaging in citizen science. This panel will provide examples from ongoing research studies demonstrating the potential for informatics to target access barriers. We will discuss innovative approaches to participant recruitment, data collection, participant engagement and communication platforms. Case studies include participants from diverse groups (across the life-span, in urban and rural settings, within the US and abroad).

Panel Description

Citizen science refers to the collection and analysis of data by various population groups in real-world settings as part of a collaboration with scientists characterized by reciprocity and mutual respect. Citizen science aims to respect and facilitate contributions from community members to scientific inquiry, build capacity, facilitate community-driven questions and research design and promote dissemination to maximize impact and uptake. The advancement of wearable and passive monitoring technologies and direct-to-consumer genetic test kits and the continued diffusion of social media have led to an increase in citizen science projects in health care. The use of platforms like Quantified Self or DIYgenomics enable individuals to participate in larger studies or engage in self-experimentation. Informatics tools that facilitate patient generated health data introduce opportunities for large cohort research studies to understand disease and drug response and the role of environmental parameters (such as living conditions, air and water quality, the built environment and available community resources) outside of clinical settings or formal organizational parameters. Citizen science and crowd-sourced health research studies are considered a potential valuable extension to traditional health and biomedical research.

The ability for individual citizens to participate in these meaningful opportunities is impacted by inequities in existing health systems. Vulnerable populations, namely communities with limited access to resources or at a higher risk to experience bias or discrimination, are less likely to engage in health related citizen science initiatives. Examples of vulnerable populations in this context include individuals and communities with limited access to digital tools that are often necessary for communication and data collection in citizen science projects, those with lack of trust in formal systems of care and academic research, or likely to experience bias, individuals with limited English proficiency or health literacy.

Informatics tools and strategies have the potential to address some of the challenges that various vulnerable populations experience in engaging in citizen science projects. This panel will provide examples from ongoing research studies demonstrating the potential for informatics systems to target access barriers. We will discuss innovative approaches to a) participant recruitment (such as crowdsourcing to understand priorities and preferences for older adults experiencing transitions in their care settings and needs); b) data collection (such as sensor-based, integrated health monitoring systems for measuring environmental, physiological, and behavioral factors in...
epidemiological studies of pediatric asthma); c) participant engagement (such as mhealth tools to interact with study participants in real time and validate behavioral choices); and d) communication platforms (such as tools to reach out to study participants in challenging circumstances like end of life care settings). The panel presentations describe case studies and examples that cover participants from diverse population groups (across the life-span, in urban and rural settings and within the US and abroad) who may be experiencing barriers to accessing health resources including accessibility, distance and bias. Furthermore, these presentations describe different technologies and study design approaches. The intended audience includes among others, informatics researchers in the areas of citizen science, and health disparities, with an emphasis on participatory research, as well as system designers and policy makers.

Participants and audience will discuss crucial questions such as:
- How can informatics tools incorporate features in their design that address digital divide and data literacy?
- What are the ethical concerns of citizen science projects for vulnerable populations?
- How can inclusive design inform the creation of citizen science tools?
- What are the opportunities and practical challenges in utilizing tools such as momentary time sampling, crowdsourcing or sensor based environmental assessment to facilitate data generated by participants in diverse settings?

The four panelists will cover the following areas:

Use of informatics tools to engage caregivers in end of life care settings (Demiris)

Family caregivers play an essential role in the delivery of care for a patient with advanced illness, especially in hospice and palliative care settings. During that time, caregivers face many physical and emotional challenges and have to coordinate many tasks and responsibilities, often with limited or no time for self-care. While interventions to support family caregivers are greatly needed, participating in research studies during that time can be very challenging. There are barriers to conducting research in the context of palliative and hospice care, and these are exacerbated for rural populations. We describe efforts to use informatics platforms to tailor recruitment and engagement of hospice caregivers within an NIH-funded randomized clinical trial (Grant Nr. R01NR012213) examining the role of technology in supporting problem solving for hospice caregivers. We discuss ethical and practical challenges in recruiting and retaining study participants in a study that is co-designed by family caregivers, and highlight strategies to maximize participant engagement and tailoring intervention material to improve acceptance and utility. We also discuss dissemination strategies to promote research opportunities for a domain that is less explored in the context of research, namely end of life care, in rural communities.

Innovative approaches for engaging difficult to reach populations in informatics research and technology design (Turner)

Involvement of users in the development and design of health information tools is a tenant of health informatics. Older adults make up a substantial portion of the US population and an even larger portion of health care expenditure and activates. While it is important to get the perspectives of a diverse group of older adults, recruitment may be difficult because of physical and social isolation and limited mobility. The challenge of limited participation and inclusion of older adults in research studies is well documented. We will discuss current efforts underway at the University of Washington to recruit and retaining older adults from a variety of settings to participate in research studies. In particular we will discuss the strengths and limitations of crowdsourcing platforms, social media and community organizational networks for recruiting difficult to reach populations. Experiences gained through several federally funded research projects such as, SOARing, SHARE-NW and Enhanced Fitness to recruit participants from rural, low income, cognitively impaired and culturally diverse populations will be shared and discussed.

Use of informatics tools to support individuals self-managing treatment (Iribarren)

Tuberculosis (TB) is an urgent global health threat and the world’s deadliest infectious disease despite it being largely curable. Poor medication adherence to TB regimens, along with challenges in monitoring patients and returning them to treatment, are important contributing factors to poor outcomes and the development of drug resistance. Individuals with TB face multiple barriers to good treatment adherence such as medication side effects, stigma, and lack of understanding about the disease and its treatment. Populations in resource limited settings with limited health literacy may experience additional challenges both in engaging in health interventions and also have access to participate in research studies. With advances in, and proliferation of, mobile technologies, there is substantial interest in the possible use of mobile interventions to address these challenges. We describe efforts to develop and refine TB treatment support tools, which link an app developed using user-centered design principles and a paper-based drug metabolite urine test strip modified for home use, to improve treatment outcomes (K23NR017210). We discuss...
interacting with study participants in real time and validating behavioral choices. We also discuss challenges for data gathering and research participant engagement in an international population in Argentina where self-management of treatment is the usual care and for a disease historically heaviest felt by marginalized populations.

Measuring environmental, physiological, and behavioral factors in epidemiological studies (Sward)
Lifestyle/behavior and environmental exposures are primary influences on disease state or wellness, particularly for adults and children with chronic conditions. Various factors such as literacy, health literacy and barriers to accessing care services may affect the ability of individuals and families to engage in extensive research studies that examine multiple factors such as the environment and the community in addition to information captured within a clinical setting. The NIH/NIBIB-funded Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) informatics center at Utah (1U54EB021973), with pediatric asthma as the exemplar condition, aims to develop and implement an informatics ecosystem to support integrated health monitoring systems for measuring environmental, physiological, and behavioral factors in epidemiological studies, using passive and wearable sensors. High-resolution sensor data need to be integrated with clinical, research-specific, and self-report data. We discuss the high-engagement participatory design process used to understand family and researcher priorities and preferences; concerns and actions to protect privacy and data security; and family’s interactions as part of the research process. Citizen science was evidenced in part by a nearly universal tendency to self-experiment, when provided real-time ability to see the data from sensors.

Participants (All panelists have agreed to participate)

George Demiris, PhD, FACMI, is a Penn Integrates Knowledge University Professor in the School of Nursing and Department of Biostatistics, Epidemiology and Informatics in the Perelman School of Medicine, University of Pennsylvania. His research focuses on the design and evaluation of patient and family centered tools to support aging in place and improve the quality of home and hospice care. He is co-founder of the Hospice Caregiving Research Network and a Fellow of the Gerontological Society of America.

Anne M. Turner, MD, MLIS, MPH, FACMI, is Professor in the Department of Health Services, School of Public Health, and Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington. Her research focuses on the application of information technology and human centered design to address public health issues involving health inequities resulting from language incongruence, aging and those living in rural environments. Dr. Turner is Associate Director of the Health Promotion Research Center at the University of Washington.

Sarah Iribarren, RN, PhD, is an Assistant Professor in Biobehavioral Nursing and Health Informatics at the University of Washington, School of Nursing. Her research focuses on the design and evaluation of patient centered innovative mobile solutions which may be key to a fairer distribution of scarce healthcare resources and a method to extend the reach of nurses and nursing to vulnerable people who may be isolated from information and healthcare services either by geography or by social or economic status.

Katherine Sward, RN, PhD, FAAN, is an Associate Professor in the University of Utah College of Nursing, and adjunct Associate Professor in the Department of Biomedical Informatics. Her research focuses on clinical and translational research informatics and decision support. She leads informatics efforts for the NICHD-funded Collaborative Pediatric Critical Care Research Network (CPCCRN) data coordinating center, and for the NCATS-funded Trial Innovation Network (TIN). She and her PRISMS collaborators are leading efforts in environmental informatics (exposure health informatics) at the University of Utah.
Reducing Burden: Evidence-based Solutions for Improving Clinicians’ EHR Experiences

Chris Dymek, EdD¹, Thomas H. Payne, MD², Genevieve B. Melton, MD, PhD³, & Hardeep Singh, MD, MPH⁴

¹Agency for Healthcare Research and Quality, Rockville, MD; ²University of Washington, Seattle, WA; ³University of Minnesota, Minneapolis, MN; ⁴Michael E. DeBakey Veterans Affairs Medical Center and Baylor College of Medicine, Houston, TX

Abstract

Electronic Health Record (EHR) use has been increasingly associated with clinician burnout. Up to 49% of a clinician’s time is spent working on EHR and other administrative activities. Given competing pressures of being clinically productive and trying to maximize face-to-face time with patients using imperfect and sub-optimized EHR systems, there is a tremendous need to fundamentally transform these systems and improve clinicians’ experiences. This panel will feature findings and implications from important health information technology (IT) research projects aimed to improve clinicians’ EHR experiences. Specifically, presentations and discussions will revolve around evidence-based solutions to improve three of the highest contributors to EHR burden: (1) documentation, (2) chart review, and (3) inbox tasks. The panel will conclude by moderating an in-depth discussion with the panel members and audience about health IT research priorities to further build the evidence-base for alleviating EHR-related clinician burden. The learning objectives for this panel include: increasing awareness of new health IT research aimed to reduce clinician EHR burden; understanding the potential for leveraging these health IT tools and applications into various healthcare settings; and understanding future health IT research directions aimed to improve the clinician EHR experience.

Panel Description

More than half of clinicians and healthcare workers in the United States experience burnout symptoms. Despite the potential benefits of EHRs, one factor consistently associated with burnout is the increasing amount of time and effort clinicians spend on EHRs. Heavy EHR usage has been associated with increased clinician burden that normally leads to losses in productivity, job dissatisfaction, increased health care costs, and even increased rates of suicide. A recent study in four different specialty practices found that 27% of a clinician’s time was spent in clinical face-to-face time while 49% was spent using the EHR and addressing other deskwork.

In order to address the EHR burden issue in a meaningful way, it is useful to understand specific areas of greatest burden. Arndt et al developed a taxonomy to help specify the percentages of time spent by clinicians in key EHR functions. The three main areas needing attention from a time perspective were documentation, chart review, and inbox tasks. In this panel, we discuss the nature of these burden areas; present current evidence for proposed solutions; and discuss promising health IT research directions. Panelists will share three use cases aimed to improve the usability and functionality of EHR systems and reduce clinician burden. The presentation of use cases will be followed by a discussant-led conversation, intended to stimulate thoughts and suggestions on future health IT research needed to build the evidence-base that will enhance clinicians’ EHR experiences. Clinical practice and policy implications will also be noted where applicable. The panelists’ work has been supported by the Agency for Healthcare Research and Quality (AHRQ), and they were selected based on the diversity of their studies related to EHR burden areas. The
The panelists represent different institutions and have specific, relevant expertise given their diverse background and training.

**Chris Dymek, EdD** (moderator and organizer) will kick off the panel by introducing the panelists and the topic including its greater significance and context in today’s healthcare landscape. Dr. Dymek will also describe the rationale and motivation for bringing this panel together.

**Thomas Payne, MD** (panelist) will present the panel’s first use case. Many clinicians are frustrated due to increased documentation requirements and EHR usability issues that result in exceedingly long timeframes to document. Recent evidence indicates that some clinicians might benefit from voice recognition software that helps create text from voice commands. Dr. Payne will present findings and implications from a research project that developed a voice-generated enhanced electronic note system (VGEENS) aimed to address problems associated with electronic progress note documentation. VGEENS has been integrated with commercial EHRs and Dr. Payne will provide insights on how this tool might provide an alternative for clinicians who prefer voice to typing to create notes.

**Genevieve Melton, MD, PhD** (panelist) will present the panel’s second use case. Chart review has always been a time-consuming task. Reviewing redundant information increases cognitive burden and causes increased risk of errors in diagnosis and treatment. Dr. Melton will present an innovative research study that applied probabilistic language modeling methods and improved automated methods to detect and visualize new, non-redundant information in patient note and reduce chart review burden. She will also present findings around in-chart reading/retrieval styles based on clinicians’ preferences and workflow demands. Differences in reviewing styles compared to the traditional ‘S-O-A-P’ format shows potential for improved EHR clinical note interfaces that need to be more aligned with clinicians’ mental models, task needs, and workflow demands.

**Hardeep Singh, MD, MPH** (panelist) will present findings and implications from a portfolio of research studies. Handling inbox tasks is an emerging burden for clinicians due to increases in new EHR-based notification types. Dr. Singh will present findings from a portfolio of research studies focused on defining, measuring and reducing clinician inbox notifications. These studies include feedback from practicing end-user clinicians and provide robust evidence needed to improve EHR inbox content and design as well as organizational workflows and policies. In addition to discussing current sociotechnical challenges related to addressing the burden of EHR inbox notifications, he will also discuss strategies to help filter messages relevant to high-quality care, EHR design examples that support team-based care, and staffing model frameworks that can assist clinicians in managing this influx of information.

**Chris Dymek, EdD** (discussant). Following presentations of the above-mentioned use cases, Dr. Dymek will lead the panel (and the audience) in a conversation about research opportunities regarding EHR burden, with a particular focus on emerging research areas of interest such as artificial intelligence (AI) including new paradigms around automation and ‘autonomation’ of clinical tasks. Questions to be considered include:

- What type of EHR-burden-related research will be most useful for clinicians and other healthcare staff? Discussion themes will address priorities as well as methods and study designs.

- What are some short- and long-term practice and policy directions to stimulate health IT research and implementation activities to both measure as well as reduce EHR-burden?

- How can EHR vendors, healthcare system leaders, and policy makers play a shared role in helping organizations make evidence-based tools/applications available and easy to use?
**Topic Rationale:**

Based on provisions in the 21st Century Cures Act, federal policy for burden reduction related to the use of EHRs and other health IT has progressed considerably. However, there remains a lack of robust evidence to inform progress in both scientific and practice transformation needed to reduce EHR and health IT burden and improve clinician experiences. This timely panel presentation will highlight three use cases and a discussion around evidence-based EHR tools and applications aimed at reducing clinician burden.

**Participation Statement:**

All proposed panelists are aware of this panel submission, and have agreed to participate in the panel if the proposal is accepted.

**Panel Participant Details:**

- **Chris Dymek, EdD.** Director, Health IT Division, Center for Evidence and Practice Improvement, Agency for Healthcare Research and Quality
- **Thomas Payne, MD.** Professor, Departments of Health Services and Biomedical Informatics & Medical Education / Medical Director, Information Technology Services, University of Washington Medicine
- **Genevieve Melton, MD, PhD.** Professor, Department of Surgery, University of Minnesota Medical Center / Chief Data and Health Informatics Officer, Fairview Health Systems
- **Hardeep Singh, MD, MPH.** Professor, Department of Medicine, Baylor College of Medicine / Chief, Health Policy, Quality & Informatics Program, Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey VA Medical Center
QDM to QUICK – Mapping the Future

Floyd Eisenberg, MD, MPH¹, Claude Nanjo, MPH², Juliet Rubini, RN-BC, MSN, MSIS³, Joe Kunisch, PhD, RN-BC⁴, Moderator: Kathy Lesh, PhD, RN-BC, CPHQ⁵
¹ESAC, Inc., Rockville, MD, ²University of Utah, Salt Lake City, UT, ³Mathematica, Richboro, PA, ⁴Memorial-Hermann Healthcare System, Houston, TX, ⁵Battelle, Arlington, VA.

Abstract

The panel will discuss how existing efforts to express electronic clinical quality measures (eCQMs) have been mapped to new, evolving standards to enable and encourage continuity with emerging health information technology (IT) efforts without disrupting existing efforts to evaluate clinical care in current healthcare settings. The work is based on efforts to harmonize standards for quality measurement and clinical decision support (CDS) using the Health Level Seven International (HL7) logical representation, Quality Improvement Clinical Knowledge (QUICK) data model, and HL7 Fast Healthcare Interoperability Resources (FHIR) Clinical Reasoning resources as the structural components and HL7 Clinical Quality Language (CQL) as the expression language. By the end of the presentations the attendees will:

☐ Understand the issues in defining data models and traversing concepts between data models.

☐ Learn challenges aligning requirements for retrospective performance measurement with existing methods by which data are captured during routine care delivery using clinical software.

☐ Learn the process for translating measurement concepts in published eCQMs to newly emerging CDS standards to provide consistent information retrieval, improve workflow, reduce burden, and increase performance rates.

☐ Describe current challenges with mapping existing eCQM data elements to existing EHR data elements to accurately capture the quality of care being delivered.

General Description of the panel and the issue(s) that will be examined and brief description of each panelist’s presentation.

The panel will discuss how existing efforts to express electronic clinical quality measures (eCQMs) have been mapped to new, evolving standards to enable and encourage continuity with emerging health information technology (IT) efforts without disrupting existing efforts to evaluate clinical care in current healthcare settings. The work is based on efforts to harmonize standards for quality measurement and clinical decision support (CDS) using the Health Level Seven International (HL7) logical representation, Quality Improvement Clinical Knowledge (QUICK) data model, and HL7 Fast Healthcare Interoperability Resources (FHIR) Clinical Reasoning resources as the structural components and HL7 Clinical Quality Language (CQL) as the expression language.

Kathy Lesh will introduce the panel and provide a high-level introduction of eCQM standards. She will also facilitate the questions and discussion.

Floyd Eisenberg will provide an overview of the Quality Data Model (QDM), a conceptual data model used to express eCQMs and discuss its evolution to align with FHIR.

Claude Nanjo will present the QUICK model, how its structure correlates to the QDM, supporting eCQM developers who might be transitioning from QDM to FHIR. He will also discuss the value proposition behind QUICK and how it enables the authoring of FHIR version-agnostic knowledge artifacts.

Juliet Rubini will discuss how measure developers have been engaged in the development and use of QUICK, showing examples of QDM-based eCQMs using QUICK and the current status of testing.

Joe Kunisch will discuss how the QUICK model aligns with existing data capture and availability within clinical settings and why alignment of the models helps to reduce clinical documentation burden. He will also discuss implementation challenges and possible mitigation strategies.

Several Centers for Medicare & Medicaid Services (CMS) quality reporting and value-based purchasing programs include eCQMs. eCQM development uses the QDM, a conceptual data model, for the core concepts that eCQM developers use to define data elements when authoring their measures by taking into consideration clinical workflow and information that should be expected within existing electronic health records (EHRs). Efforts by HL7, standards development organization, to harmonize data models for clinical quality measurement (CQM) and CDS developed QUICK, a logical data model. QUICK was developed to enable eCQMs to reference required information while allowing local implementation sites latitude to use HL7’s FHIR version(s) they have available without requiring different versions of an eCQM. QUICK, is an implementation guide (IG) to help eCQM and CDS artifact developers create and test measure expressions for data they should be able to expect in existing software implementations using FHIR for eCQMs and CDS. FHIR’s core maturity model requires evidence of data
availability in 80% of existing clinical software. Thus, QUICK represents a surrogate method to evaluate clinical information without disrupting existing clinical workflow and data capture while allowing each implementation site to use the version of FHIR in their existing software. This current effort to translate QDM to QUICK enables transformation of eCQM queries to consistently retrieve information, improve workflow, reduce burden, increase performance rates, and share logic with CDS rules, thereby enhancing clinical quality prospectively.

Figure 1 details efforts in HL7 to evolve standards for eCQMs over time defining the structure, the data model, and the expression language.

The QDM version for eCQM reporting in 2021 (version 5.5) specifically addresses concepts as they are specified in the HL7 FHIR IG known as QUICK. QUICK is a logical model with an IG for using FHIR to author quality measures and CDS artifacts with the FHIR Clinical Reasoning resource (e.g., measure, measure report, Event-Condition-Action-Rules). The relationship between QDM version 5.5 and FHIR is significant as the maturity model for the base FHIR resources requires that they are available in 80% of existing EHRs; hence, the newest versions of QDM should represent concepts present in current clinical information systems. Subsequent QDM versions (version 5.5 planned for measure reporting in 2021) will also be mapped to QUICK. Some existing eCQMs expressed in QDM have been converted to FHIR consistent with the QUICK IG using fabricated data in the HL7 FHIR Connectathons. CDS artifacts have been created based on a few eCQMs and tested as well. Floyd Eisenberg will discuss some of the challenges mapping information required for defining concepts (i.e., the QDM) with existing FHIR resources, and how those challenges have been mitigated through the harmonization effort. Joe Kunisch will discuss how the use of a logical data model can facilitate coordinated eCQM and CDS development and implementation.

The QUICK Update (version 3.0.2), published February 2019, includes mapping of QDM categories of information and attributes to the QUICK concepts and metadata. The anticipated September 2019 HL7 balloted version of QUICK will be FHIR version neutral such that users can specify the FHIR version implemented at each site to retrieve information for eCQMs or clinical decision support (CDS) CQL expressions. The QDM to QUICK mapping is significant because QDM has evolved over the past seven to eight years to a much more expressive and detailed conceptual data model that benefits eCQM development. The mapping to QUICK allows CDS developers to more directly translate eCQM concepts to CDS artifacts to improve the capture and use of data in an EHR that supports and represents the quality of care being delivered to a patient. QUICK is predominantly based on FHIR resources with some extensions to express metadata not present in basic FHIR resources. Most of the QDM concepts map directly to FHIR resources and extensions represented in QUICK. The mapping tables provide bidirectional mapping, i.e., QUICK concepts to corresponding QDM datatype(s) and attributes, and QDM datatype and attribute mappings to QUICK resources and metadata. In some cases, multiple QDM datatype and attribute mappings to QUICK resources and metadata. In some cases, multiple QDM datatypes map to a single QUICK resource as indicated in the QUICK mapping tables. Some areas did not map well, leading to recommendations to base FHIR resources for clarification, or modifications to the QDM conceptual model. Table 1 shows some examples of the mapping exercise. Floyd Eisenberg will provide more details during the presentation. In some cases, the base FHIR resources were purposely non-specific, leading to conflicts as subsequent FHIR versions added granularity. For example, until subsequently
defined by the HL7 Workgroup managing the resource, blood pressure could be specified as two observations, a systolic and a diastolic reading, or as a single observation with two components: blood pressure, systolic component, and diastolic component (the latter was the more recent decision).

Table 1. Examples of QDM to HL7 QUICK Mapping Challenges.

<table>
<thead>
<tr>
<th>QDM Concept and Context</th>
<th>QUICK Concept</th>
<th>Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic Study, Performed Laboratory Test, Performed</td>
<td>DiagnosticReport Observation</td>
<td>When to use DiagnosticReport, Observation, and Observation.component</td>
</tr>
<tr>
<td>Assessment, Recommended Assessment, Order Procedure, Recommended Procedure, Order</td>
<td>ProcedureRequest.intent</td>
<td>ProcedureRequest.intent has several options; most QDM recommendations are mapped to intent plan, QDM order concepts are mapped to intent order.</td>
</tr>
<tr>
<td>(and other QDM concepts addressing recommendations Vs. orders)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All panelists will discuss their role in the process employed to complete the mapping effort, comments received during the ballot process, and resolution prepared for a September 2018 HL7 ballot to align the CMS QDM effort with QUICK, allowing direct translation of existing measures to FHIR, and enabling translation of existing measures to clinical decision support artifacts using FHIR as well.

An explanation why the topic of this panel is timely, urgent, needed, or attention grabbing is required with a discussion of anticipated audience.

eCQMs continue to use the QDM and, beginning with the 2019 reporting/performance period, use CQL as the expression language. The expression of eCQMs using QDM as the logic expression presented many challenges for EHR vendors and healthcare organizations to construct and deploy CDS and eCQMs due to the logic and constraints of the QDM. This created significant barriers for healthcare organizations in adopting and successfully implementing eCQMs that captured an accurate portrayal of the quality of care provided to a patient. Using CQL as the logic expression language has alleviated many of those barriers. Currently, many organizations and vendors are evaluating HL7 FHIR and developing clinical decision support processes to improve quality in real-time. The CDS and eCQM efforts can best align if mapping exists to translate the eCQM content to information required to manage CDS. Broad-based programs will not quickly change to new, emerging standards and consideration of such change will require evidence that the health IT industry has tested and is using the new standards routinely. Therefore, a map from the existing eCQM data model to the new and emerging standard, QUICK, provides a mechanism to work with existing and innovative approaches to improving quality with health IT. Tooling to automate the mapping may become available to enable direct translation of a QDM-based eCQM to QUICK/FHIR-based measurement and CDS.

Discussion questions to enhance audience participation.

- What challenges might you expect if quality measures are defined using different descriptions of data than those used to capture information during routine clinical care?
- How will the ability to translate information in clinical quality measures to match the way information exists in clinical software impact you?
- What is the benefit of clear, unambiguous definitions in determining a clinician’s level of performance with respect to a clinical guideline?
- What is the likelihood of reducing clinical documentation burden by assuring that clinical quality measures and clinical decision support efforts define information the same way?

All participants have agreed to take part on the panel.
Learning Across a Healthcare Data Network to Improve Model Robustness and Evidence Reliability

Noémie Elhadad, PhD¹, Jenna Reps, PhD², Alison Callahan, PhD³, Iñigo Urteaga, PhD¹, Patrick Ryan, PhD²
¹Columbia University, New York, NY; ²Janssen Research and Development, Titusville, NJ; ³Stanford University, Stanford, CA;

Abstract
Recent work in machine learning and modeling from healthcare data has shown much promise in leveraging large datasets to infer signal and predict outcomes of interest. With increasing numbers of such data sources available via collaborative efforts such as those created by the Observational Health Data Science and Informatics (OHDSI) network, it is now possible conduct clinical studies at an unprecedented scale. In this didactic panel, we will focus on another contribution of such data networks beyond the increased power of larger sample size, namely learning across multiple data sources as a way to improve the robustness of the learned models and increase the reliability of derived evidence. Our panel participants are junior and mid-career researchers in the field of machine learning for healthcare. We will (1) review the OHDSI network as one example of such a data network, (2) describe use cases in predictive analytics and an on-demand informatics consult service to distill clinical evidence from such data networks, where increased reliability is achieved through learning across different types of observational data, (3) describe current efforts in facilitating external model validation in OHDSI, and (4) illustrate current efforts in machine learning and specifically transfer learning to increase model robustness.

Learning Objectives
- Understand the landscape of healthcare data sources available to informatics and data science researchers, along with the OHDSI common data model across such data sources;
- Appreciate the need for model robustness and increased evidence reliability;
- Enumerate current efforts in increasing reliability of evidence through external validation across the OHDSI network;
- Learn about current efforts in machine learning for improving robustness and reliability of learned models across different datasets;

Panel Participants
Noémie Elhadad, PhD, FACMI is an Associate Professor of Biomedical Informatics at Columbia University. She is affiliated with the Columbia Computer Science Department and the Columbia Data Science Institute. Her research is in machine learning and natural language processing applied to clinical and patient-generated datasets.

Jenna Reps, PhD is a Senior Epidemiology Informaticist at Janssen research and Development. Her work is on novel solutions to personalize risk prediction and is an active contributor in the OHDSI tool suite. She has applied her work on a variety of healthcare prediction problems.

Alison Callahan, PhD is a Research Scientist at the Center for Biomedical Informatics at Stanford University. Her work involves research and development of informatics methods for the analysis of biomedical and clinical data, to derive insights and inform medical decision making.

Iñigo Urteaga, PhD is an Associate Research Scientist in Applied Physics and Applied Mathematics at Columbia University and is affiliated with the Columbia Data Science Institute. His work is in machine learning and signal processing in unsupervised and supervised settings. He has applied his work to clinical and patient-generated datasets.
**Patrick Ryan**, PhD is Senior Director of Epidemiology and the Head of Epidemiology Analytics at Janssen Research and Development, where he is leading efforts to develop and apply analysis methods to better understand the real-world effects of medical products. He is an active collaborator in OHDSI. He served as a principal investigator of the Observational Medical Outcomes Partnership (OMOP), which led to OHDSI.

**Timeliness of Panel**

The fields of machine learning and data science for healthcare are witnessing a drastic increase in interest, at the same time that the amount of and access to healthcare data is rising. In clinical research, for a statistical model to be used effectively it is fundamental that there is extensive understanding of the circumstances where it is reliable. The performance of a model on a single dataset is often insufficient to assess its broad applicability, and external validation is required to generate evidence supporting the model’s performance in similar or different patient populations. Reviews have shown there is a lack of external validation in clinical research [1,2] and this is likely because replicating a model in new data is a challenging task. Large healthcare data networks present an unprecedented opportunity to externally validate statistical models learned on one source by evaluating it against other sources. This capability makes it possible of testing the transportability of models across a data network. The lack of reliable evidence is also problematic at the point of care. Most medical decisions are made without the support of rigorous evidence [3,4]. When treatment guidelines do exist, they often do not apply to complex patients commonly seen in the clinic. As a result, clinicians must often rely exclusively on their own experiences and that of their colleagues when making treatment decisions for the patient in front of them. In contrast, the availability of data and effective methods to analyze it have been transformative in other disciplines, from information technology to law. There is an unprecedented opportunity to realize this same transformation in healthcare [4]. The concepts and ideas discussed in this panel—augmenting data networks with principled approaches to model learning and novel efforts in transfer learning—can make this opportunity a reality.

**Panel Description**

Dr. Elhadad will be the moderator of the panel. She will motivate the need for this didactic panel, and present on the Observational Health Data Sciences and Informatics (OHDSI) multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics as an example of a health data network [5]. She will then present a use case for increased model reliability when learning across data sources in the context of predictive analytics. Specifically, she will show how models that learn from claims data increase the power of study and can be leveraged and validated against electronic health record (EHR) datasets to create early detection of disease decision support modules.

Dr. Repps will present on the power of external validation across the OHDSI health data network. She will present tools and methods currently deployed in the OHDSI tool suite to facilitate the external validation of predictive models. She will showcase how the OHDSI standardizations have made sharing models between sites very easy and quick. This capability creates the possible of testing the transportability of models across a data network [5].

Dr. Callahan will present on methods developed for and learnings from offering an informatics consult service [6] to Stanford Health Care clinicians, to rapidly distill evidence from EHR data to inform clinical decisions. She will describe the consult service, including strategies to leverage multiple sources of clinical evidence, and the consult team’s experience using the service as a test bed to evaluate the robustness and reliability of statistical methods developed for analyzing observational health data.

Dr. Urteaga will present machine learning methods for domain adaptation, and how they can be applied to EHR data networks. Transfer learning has become a popular area of machine learning whereby knowledge gained from one source can be integrated with knowledge gained from another source to derive a 'repurposed' composite model that exhibits optimal operating characteristics. This technique opens interesting opportunities if applied across an EHR data network. Dr. Urteaga will present recent work on high-throughput joint unsupervised learning of phenotypes across different EHRs and different clinical settings (e.g., acute care, intensive care, and outpatient care).

Dr. Ryan will present work on improving confidence in causal inference through network analysis. Observational studies are challenged in their ability to produce reliable population-level effect estimation by threats of bias due to confounding, selection bias, and measurement error. Bradford-Hill’s causal viewpoints identify multiple dimensions
beyond strength of association, which can be used to augment the argument of causal effect. This includes 'consistency', which can be operationalized by replication of observed effects across multiple disparate data sources. His presentation will highlight the incremental value of study replication across a data network to increase the confidence in the causal assessment of exposures and outcomes.

**Target Audience**

There are several audience members who might find this panel of relevance: (1) clinical researchers interested in deriving reliable knowledge from large healthcare datasets; (2) data scientists and practitioners who experiment with statistical learning models; (3) informatics and data science trainees interested in learning more about current efforts in statistical methods and machine learning for healthcare.

**Statement of Participation**

All panel participants have agreed to participate in the panel and have contributed to this proposal.

**References**


Sync for Genes: Integrating Genetic Information at the Point of Care

Stephanie Garcia, MPH¹, Robert R. Freimuth, PhD², David E. Jones, PhD³, Bob Milius, PhD⁴

¹Office of the National Coordinator for Health Information Technology, Washington, DC; ²Mayo Clinic, Rochester, MN; ³Utah Department of Health, Salt Lake City, UT; ⁴National Marrow Donor Program, Minneapolis, MN

Abstract

Genomic data sharing is critical to the Precision Medicine Initiative (PMI) and the next impetus of genetic research. Launched in 2016, Sync for Genes aims to standardize the sharing of genomic information between laboratories, providers, patients, and researchers. Two phases of this project have culminated, both resulting in the advancement of the Health Level 7 International (HL7®) Fast Healthcare Interoperability Resources (FHIR®) Clinical Genomic profile. During the course of this project, several important gaps were identified in the emerging standard that must be filled to set a firm foundation for the use of genomic data in the provision of tailored medicine and research. This panel will share results and findings from the Sync for Genes project, review the semantic data model work conducted to date, and discuss with the audience priorities for future work. The discussion will explore roadblocks that may be hindering widespread adoption of FHIR for genomics and weigh alternatives that may be in use to support the exchange of genomic data for use at the point of care and for research.

Introduction

Precision medicine is an innovative approach for disease treatment and prevention that takes into account individual variability in lifestyle, environment, and biological makeup. It acknowledges that each person is unique and that the interaction of these factors greatly impacts our health. Delivering on the promise of precision medicine will require a variety of data types to conduct precision medicine research and generate evidence for more effective prevention, diagnosis, and treatment. This information must also be available for the effective delivery of tailored treatment.

While work on collecting and integrating lifestyle and environment data is still emerging, the amount of genetic information available has greatly increased, in part due to the reduced costs of genetic testing and increased understanding of the role genetics plays in our lives. This is particularly true in pharmacogenetics and oncology. Effective use of these results requires that they be shared between laboratories, providers, and with patients. Today, most genetic test results are shared in human-readable, but not machine-readable form. In addition, the organization, content, and formatting of results from different testing laboratories are often inconsistent, making it difficult to integrate genetic test results into electronic health record systems and utilize them effectively at the point of care.

More recently, the All of Us Research program under the Precision Medicine Initiative will collect an unprecedented number of biosamples and other relevant health data (including genetic information when available) to enable precision medicine research. This creates a sense of urgency for ensuring genetic test results can be shared in standardized forms both for conducting the research and sharing genetic sequencing results with participants.

Sync for Genes

In 2016, the Office of the National Coordinator for Health Information Technology (ONC) launched the Sync for Genes project in collaboration with the National Institutes of Health (NIH) with the goal of standardizing the sharing of genomic information between laboratories, providers, patients, and researchers. For the first phase of this project, ONC partnered with the Food and Drug Administration (FDA) and leading private sector organizations to take the first step toward integrating clinical genomics into the clinical point of care by expediting the use of standards, such as Health Level 7 International (HL7®) Fast Healthcare Interoperability Resources (FHIR®). The first phase focused on testing and validation of FHIR resources and included use cases from five pilot sites. Those results have been compiled into a final report published in 2018 and helped inform the development of FHIR genomics STU3.

The second phase of Sync for Genes, launched in 2018, took the next step from use cases (explored in Phase 1) to actual implementation. Four pilot sites expanded the profile work to demonstrate proof-of-concepts that demonstrated the exchange of genomic data in different clinical workflows. The goals of the pilot projects for Phase 2 included:
1. Utah Department of Health Newborn Screening: Share data with providers in real-time for the care of at-risk and vulnerable newborns
2. Lehigh Valley Health Network (LVHN): Pharmacogenomics: Leverage genomic content for diagnosis, care management, & pharmacogenomics to support their advanced cardiology program
3. National Marrow Donor Program (NMDP): Patient/Donor Matching: Convert existing Next Generation Sequencing (NGS)-based human leukocyte antigen HLA genotyping information to FHIR without losing completeness of the data
4. Weill Cornell Medicine: Cancer Genomic Decision Support: Convert discrete genomic results to FHIR and pair data with a clinician-facing app that provides supplemental knowledge of variant results

The Phase 2 pilot sites successfully demonstrated the exchange of a pilot site-specific FHIR message containing genomic data using a server preconfigured with FHIR STU3 resources and profiles. The initial findings from this project, have been shared with the HL7 Clinical Genomics and Orders & Observations work groups to continue to inform the development of the corresponding FHIR specifications. The Phase 2 pilot sites also uncovered challenges with genomic diagnostic reports in general and issues that have industry-wide implications (Table 1). While some of these challenges are outside the scope of this project, they have implications for both research and clinical care. A full report of the outcomes from the Sync for Genes Phase 2 project will be released in summer 2019.

Table 1. Challenges and Gaps

<table>
<thead>
<tr>
<th>Genomic Testing Reports</th>
<th>Industry-Wide Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Do not always contain the same information (variability in data content and granularity)</td>
<td>• Data modeling and semantics</td>
</tr>
<tr>
<td>• Different use cases have unique data needs (variability in data elements)</td>
<td>• Data provenance</td>
</tr>
<tr>
<td>• Research use cases require the exchange of data files such as variant cell format (VCF) in addition to the clinical report</td>
<td>• Privacy and security of sensitive data</td>
</tr>
</tbody>
</table>

Given the findings of the first two phases, Sync for Genes has shifted its focus to the development of robust data models to guide implementers by providing information regarding the semantics of data elements. This will include the definition, structure, conditions or constraints under which data are collected, and relevant value sets, as well as how those data elements are related to one another. This will facilitate implementation, enable both more effective sharing of genetic information, and assist with transition from one version of FHIR to the next.

Panel Objectives and Presenters

The aim of this panel is to share the results and findings from the Sync for Genes project, discuss semantic data model work conducted to date, and discuss with the audience priorities for future work. The panel brings together the informatics and research experts leading this work at ONC, Mayo Clinic, Utah Department of Health, and the National Marrow Donor Program.

Stephanie Garcia, MPH (moderator and organizer) oversees the Sync for Genes project as well as the ONC PCOR portfolio, which is a suite of projects that invests in the health information technology infrastructure that supports the health research enterprise. She has collaborated with public and private stakeholders, successfully executing projects that advance the effective use of health information technology to improve the nation’s health. Ms. Garcia’s will introduce the session, provide an overview of the Sync for Genes project, and moderate the session.

Dr. Robert Freimuth leads a variety of research studies and standards development activities focusing on genomics, including Sync for Genes, HL7, and the Global Alliance for Genomics and Health (GA4GH). He will discuss overall Sync for Genes phase 2 project findings and will present an overview of the semantic data model work that is being done within HL7 Clinical Genomics.

Dr. David Jones is a Senior Health Informaticist at the Utah Newborn Screening Program. He will discuss his team’s experience in exploring the use of FHIR genomics to facilitate the exchange of genomic data within the field of newborn screening. His perspective will also touch upon considerations particular to state organizations and working with data sharing partners.
Dr. Bob Milius is a Principal Research Scientist at the National Marrow Donor Program (NMDP). He will discuss his team’s experience with using FHIR genomics to facilitate the exchange of genomic data by converting existing next generation sequencing (NGS) based human leukocyte antigen HLA genotyping information to FHIR without losing completeness of the data. His perspective will also touch upon his experience with implementing genomics for other use cases relevant to NMDP and working with a standards development organization to advance the technology for the exchange of genomic data.

Panel Discussion Questions

- What is the current status of the HL7 FHIR Genomics implementation guide?
- What were the biggest challenges pilot sites faced when implementing FHIR genomics?
- What are some of the steps that your organization took to address those challenges and what would you recommend to others implementing FHIR genomics?
- How will the data modeling work inform the development of genomics standards?
- What additional informatics work is needed to enable effective integration of genomic information at the point of care?
- What issues has the audience faced when trying to implement FHIR genomics or exchange genomic data?
- What are the highest priorities for future work?

Panel Learning Objectives

1. Participants will have a better understanding of the goals of the Sync for Genes project and findings to date.
2. Participants will learn about the testing conducted by Sync for Genes pilot sites and the implications for standardized sharing of genomic information.
3. Participants will learn about key gaps that exist in the current specification and how they are being addressed.

Conclusion

Delivering precision medicine in the era of genetic testing and sequencing will require effective and efficient sharing of genetic information. Sync for Genes is enabling the sharing of standardized genomic information between laboratories, providers, patients, and researchers as the first path in making genetic data available at the point of care. This panel will share results from this project and gather participant input to inform future work at ONC and NIH.

Statement of Agreement to Participate

All panelists have approved of this submission and agreed to participate.

References

Promise, challenges, and risks of digital data in mental health research and care

Sharath C. Guntuku, PhD¹, Lauren Southwick, MPH¹, Ian Barnett, PhD¹, Ipsit Vahia, MD², Stevie Chancellor, PhD³

¹University of Pennsylvania, Philadelphia, PA, USA  
²Harvard Medical School, Boston, MA, USA  
³Northwestern University, Evanston, IL, USA

Abstract

Digital data available from social media and passive sensing offer a new window into individual health, with the potential to uncover a nuanced understanding of an individual’s lived experiences, both expressed and unexpressed. Harvesting health insights from digital data is significant yet understudied with implications to advance efforts that capitalize on objective digital data and evidence to improve health forecasting and interventions. In this panel, we will explore how digital data can be used to characterize, forecast mental health conditions, and potentially use it in clinical settings. In the first talk, we discuss how digital data may impact the process of care, specifically the doctor-patient relationship. The second talk explores the potential of real-time detection of behavioral anomalies using passive sensing to forecast schizophrenia. In the third talk, we show how social media can be used to better understand the naturalistic manifestations and consequences of loneliness based on a cohort of Twitter users. The fourth talk examines emergent ethical and methods challenges of utilizing novel digital data sources in mental health research and care. Together, this panel demonstrates the potential for social media and wearable sensors to uncover insight into psychological health and well-being. This panel is designed to explain a new generation of innovative and targeted digital decision support tools, and launch large-scale trials exploring the acceptability, efficacy, and ethical as well as clinical implications of incorporating these tools into clinical care.

Introduction

When individuals post to social media, or use wearable devices, digital data generated through these everyday interactions with technology reveal a great deal about individuals’ behaviors in ways that were previously not observable. Prior work offers particular promise in being able to identify early signals from digital data about individuals’ emotions and mental health (1,2).

The speed with which data is generated creates potential for earlier signal about disease development or exacerbation. The regular production of data also allows for tracking of health in real time or over time, using insights not available from traditional surveys or insurance claims (3). Digital platforms not only can provide markers of health but also serve as platforms that can be used for direct intervention (4).

In this panel, we discuss the promise, challenges and risk (5) associated with the utilization of digital data in health forecasting and interventions. The first talk will focus on the current state of clinical practice using data from two clinician surveys, with the goal of capturing current clinician practices, eliciting preferences and concerns related to including data from electronic media in patient care. The second talk examines the potential of passive sensing data for personalized forecasting of psychotic relapse hospitalizations. The third talk will examine the shows how social media can be used to better understand the naturalistic manifestations and consequences of loneliness based on a cohort of Twitter users. The fourth talk will discuss emergent ethical and methodological challenges of tapping into digital data traces to understand and measure individuals health.

Incorporation of Electronic Media into Clinical Care (Vahia)

This study will discuss how digital data may impact the process of care, specifically the doctor-patient relationship. We present data from two clinician studies that capture physician attitudes and current practices in the US and India. Obtaining collateral information from a patient is an essential component of providing effective psychiatric and
psychotherapeutic care. Research indicates that patients' social and electronic media contains information relevant to their psychotherapy and clinical care. However, it remains unclear to what degree this content is being actively utilized by clinicians as a part of diagnosis or therapy. Moreover, clinicians’ attitudes around this practice have not been well characterized.

An online survey was sent to outpatient psychotherapists aimed at identifying what extent and with which patients they utilized this practice, as well as their reasons for or against doing so. Of the 115 total respondents, 62% indicated that they had viewed at least one patient’s social or electronic media as part of psychotherapy and 65 of those 71 (92%) endorsed being able to provide more effective treatment as a result of this information. Moreover, analysis of survey responses found patterns of use associated with clinicians’ years of experience and patient demographics, including age and primary diagnosis. This study will also present findings from a similar but smaller survey conducted with psychiatrists in India in February 2019.

**Forecasting Schizophrenia Relapse using Passive Sensing (Barnett)**

Among individuals diagnosed, hospitalized, and treated for schizophrenia, up to 40% of those discharged may relapse within 1 year even with appropriate treatment. Passively collected smartphone behavioral data present a scalable and at present underutilized opportunity to monitor patients in order to identify possible warning signs of relapse. Seventeen patients with schizophrenia in active treatment at a state mental health clinic in Boston used the BeeWe app on their personal smartphone for up to 3 months. By testing for changes in mobility patterns and social behavior over time as measured through smartphone use, we were able to identify statistically significant anomalies in patient behavior in the days prior to relapse. We found that the rate of behavioral anomalies detected in the 2 weeks prior to relapse was 71% higher than the rate of anomalies during other time periods.

Our findings show how passive smartphone data, data collected in the background during regular phone use without active input from the subjects, can provide an unprecedented and detailed view into patient behavior outside the clinic (6). Real-time detection of behavioral anomalies could signal the need for an intervention before an escalation of symptoms and relapse occur, therefore reducing patient suffering and reducing the cost of care.

**Measuring and Characterizing Loneliness using Social Media (Guntuku)**

Loneliness affects approximately 30% of individuals in the United States, and is associated with mental health conditions and high morbidity. We characterized the (online) lives of people who express being lonely and correlate their posts with predictors of mental health. We collected more than 400 million tweets posted by 506,216 users in Pennsylvania, USA, from 2012 to 2015. We identified users who had at least five posts containing the words ‘lonely’ or ‘alone’ (lonely group) and compared them to a control group matched by age, gender, and period of posting. Compared to the control group, users in the lonely group contained more posts about difficult interpersonal relationships, psychosomatic symptoms, substance use, wanting change, unhealthy eating, having troubles with sleep, and in general posted more at night. These posts were also associated with linguistic markers of anger, depression, and anxiety. A random forest model predicted expressions of loneliness online with an accuracy of 77% (AUC .86). Our approach can inform online surveillance for high-risk individuals experiencing loneliness and interventions focused on addressing morbidity in this condition.

**Ethical Tensions, Challenges, and Potential Solutions for Predicting Mental Health Status from Social Media (Chancellor)**

Powered by machine learning techniques, social media has been successfully employed to predict mental health states of individuals such as depression, eating disorders, and the risk of suicide. These algorithmic inferences hold great potential in supporting early detection and treatment of mental disorders and in the design of interventions. At the same time, the outcomes of this research can pose great risks to individuals, such as issues of incorrect, opaque algorithmic predictions, the involvement of bad or unaccountable actors, and potential biases from intentional or inadvertent misuse of insights.

In the panel, Chancellor will briefly discuss her findings with her work on this topic area. She will discuss the emergent ethical challenges of conducting this work. This work poses key questions around topics such as equity and fairness, privacy and engagement with stakeholders, issues of validity of methods, and potential for intervention (7), such as: How do you gain meaningful consent to conduct social media research for mental health, if consent is necessary at all? How do you unite findings from psychology and psychiatry to precisely identify behaviors in social
media? What are the changing roles for stakeholders in this process, such as clinical partners and social media users whose data is used to generate these predictions? She will also discuss new findings about these tensions from a literature survey of research conducted in this field. This literature survey draws from 55 papers across health informatics, human computer interaction, and machine learning, that predicts the status of mental health using social media data. These challenges span both methods areas in computer science as well as topical areas for ethics, privacy, clinical psychiatry, and human-centered design. She will also discuss recommendations for to begin to solve these pressing ethical challenges.

An explanation why the topic of this panel is timely, urgent, needed, or attention grabbing is required with a discussion of anticipated audience.

As digital data becomes more ubiquitous and incorporated into decision making for health, it is imperative we discuss the benefits and risks of incorporating this data into practice. Additionally, we believe this panel is well-suited and timely for AMIA given the interdisciplinary audience of the conference as well as the impact such discussions can have on establishing standards of practice within this space. Within the last several months, several articles in journals, such as Nature, New England Journal of Medicine, JAMA, and AMIA were published on this topic, some of which include works by co-authors on this panel. We believe this panel complements AMIA’s conference mission and advances multi-disciplinary discussions around a complex and evolving topic of mental health research and care.

Discussion questions to enhance audience participation

1/ Which of the following modalities of electronic communication was identified by clinicians to be most frequently accessed as part of therapy? a) Facebook; b) Text Messaging; c) Email; d) Instagram

2/ Under the EU General Data Protection Regulation (GDPR), what might be patients and providers’ view on incorporating digital data in health care?

3/ Could digital data’s incorporation in health care further health disparities among marginalized populations?

4/ Considering the role of mental health apps, how will they shape predicting, detecting, and intervening within the healthcare system?

5/ What are the challenges in ethical applications of your research to these topic areas?

References

5. Sherry Pagoto, Camille Nebeker; How scientists can take the lead in establishing ethical practices for social media research, Journal of the American Medical Informatics Association, Volume 26, Issue 4, 1 April 2019, Pages 311–313, https://doi.org/10.1093/jamia/ocy174

Note: The panel organiser (Sharath C. Guntuku) confirms that all participants have agreed to take part on the panel.
Variability of Nursing Workforce Characteristics: Impact on Outcomes

Ellen M. Harper, DNP, RN, MBA, FAAN; Amy Garcia, DNP, RN, CENP; John M. Welton, PhD, RN, FAAN; Lisa A. Moon, PhD, RN, CCMC, LNC

1University of Kansas, Kansas City, KS; 2University of Colorado, Denver, CO; 3University of Minnesota, Minneapolis, MN.

Abstract

Examining variation in clinical practice is an important step to measuring efficiency and effectiveness in care delivery. The study of practice variation by an individual nurse and patient characteristics can indicate potential opportunities to reduce costs and improve the value of healthcare delivery.

This didactic panel will describe the overall approach taken to examine the clinical practice of nurses linked to patients at a single pediatric hospital. This exemplary of emerging data science uses data from multiple electronic data sources, including clinical, administrative, human resources, staffing, and financial systems to identify outcomes with a new level of precision. Examples include nurse-sensitive outcomes throughout each shift, length of stay, and traditional discharge measures. Administrative outcomes include calculations of the direct cost of nursing care by nursing unit and diagnosis. Attendees will become familiar with the Nursing Value Data Model and methodology for working with big data. They will examine variation in nursing clinical practice and discuss their findings, including how it can be better supported for clinical and translational research applications.

Key words: Nurse Value, data science, healthcare quality, practice variation, interoperability, standards

A general description of the panel and issues that will be examined

Examining variation in clinical practice is an important step to measuring efficiency and effectiveness in care delivery. The study of individual nurses, nursing practice, and various patient characteristics can indicate potential opportunities to reduce costs and improve the value of healthcare delivery. Nurse-sensitive outcomes measured in the National Database of Nurse Quality Indicators (NDNQI) represent the concentration of nursing knowledge at the unit level, making it difficult to visualize an individual nurse’s contribution to patient outcomes. Measuring outcomes at the unit or organizational level, (e.g., nursing care hours per day), is not granular enough to reveal the structures and processes that occur within the nurse-patient dyad.

Significant progress has been made in creating terminological and structural standards for representing general clinical data. Yet, standards and interoperability solutions for nurse-sensitive clinical data have not yet achieved the same level of maturity. Standardization of data presents challenges including the lack of structured data elements in the source data systems, poor data governance practices, and a disconnect between clinical practice and research.

This didactic panel will describe the overall approach and results from a retrospective observational research study used to examine the clinical practice variation of nurses linked to patients at a single pediatric hospital. This exemplary of emerging data science uses data from multiple electronic data sources, including clinical, administrative, human resources, staffing, and financial systems to identify outcomes with a new level of precision. Clinical outcomes include progress or deterioration (acuity score) toward nurse-sensitive outcomes (Nursing Outcomes Classification (NOC) scores) throughout each shift, and length of stay (LOS) measures. Administrative outcomes include calculations of the direct cost of nursing care by nursing unit and APR-DRG diagnosis.

Importance of the panel topic

This didactic panel builds on the current national efforts of the Nursing Value Data Model (NVDM) structured methodology, organized through the Nursing Big Data to Knowledge initiative hosted at the University of Minnesota School of Nursing. The NVDM is a standardized, feasible, valid, and relevant data and technology infrastructure used to routinely collect and aggregate nursing-centric data from disparate electronic systems, including electronic health records (EHR), staffing, human resources information systems (HRIS), staff education, and financial systems. Each data element is further defined for clarity and uses definitions from established taxonomies, including the Centers for Medicare/Medicaid Services (CMS), the Nursing Management Minimum Data Set (NMMDS), and the NDNQI.

The panel will include presentations on the following four topic areas:
1. Describe the background and history of the Nursing Value Data Model (NVDM) and introduce the pilot study and how the data dictionary data collection methods used to demonstrate the relationships across disparate electronic data using primary keys (relationships), including data sources, sample, setting and time frame, and nursing care. The broad research questions examined in the study are: 1) Do the characteristics (certification, education, tenure) of nurses assigned to care for a patient significantly impact patient outcomes? 2) What was the variability of the Discomfort Level NOC scores following surgery and total LOS? Do the characteristics of the nurses assigned impact the NOC scores? 3) What are the nursing financial outcomes (e.g., labor costs), and are these influenced by patient, nurse, or unit characteristics?

2. Methods - Models for micro-costing and patient outcomes were developed from data extracted, de-identified and encrypted before secure transfer to HIPAA compliant research repository. Components are illustrated in Figure 1. Sources include data from the EHR, registered nurse (RN) - patient assignments, patient acuity, NOC scores, and characteristics of the nurse from HR and employee data.

![Figure 1](image1.png)

Figure 2
After initial data set up and cleaning, new variables were created e.g. direct nursing costs. The final dataset allows researchers to link each nurse to patients assigned and examine changes in patient’s acuity multiple times during the shift. The approach directly and identifies both patient and nurse-level data (e.g., direct costs of care for each nurse can be calculated using actual hours of care and wage). The approach provides a level of granularity about the moment-to-moment clinical trajectory of patients that have been difficult to capture to date.

3. Sample and Results. This IRB-approved study uses a descriptive, retrospective analysis of 43,936 children admitted to a pediatric hospital over a three-year period (2014 – 2016), with 1,728 nurses assigned to care for them. Qualitative analysis of repeated measures of acuity by unit type is represented in Figure 2. Using inferential analysis of all nurses’ assignments and wage we were able to total the overall RN costs per patient/shift by unit type (Figure 3).

![Figure 2](image2.png)

![Figure 3](image3.png)

Note: these are actual direct costs adjusted for wage, shift differential, and holiday. Does not include indirect costs, e.g. vacation/sick time, benefits.
4. Future directions. Data sensitive to nurses’ are underrepresented in large clinical trials and national clinical research networks. Some national networks e.g. Patient-Centered Outcomes Research Institute use a distributed data model to harness the power of data from multiple sources while taking advantage of partnerships between patients, clinicians, and health systems. The structured methodology presented provides a model for usable and actionable information that can be the basis to improve the quality and efficiency of nursing care delivery systems. Such information provides an understanding of how to optimize nurse staffing and assignment patterns that will improve outcomes by aligning individual nurse’s characteristics (experience, education, certification) to individual patient’s needs.

Discussion questions:
1. How can information in the NVDM be used to quantify and measure nursing value?
2. What nursing activity and operational aspects and efficiencies can be visualized to measure nursing value?
3. Are there better ways to staff nurses? (e.g., put the more experienced nurse on first day post-op)?
4. Is it possible to extend the NVDM methodology to increase the reliability of nurse-sensitive data sharing by creating the foundational elements required for interoperability amongst disparate healthcare systems?

Learning Objectives
1. Attendees will learn about the Nurse Value Data Model and methodology for implementation.
2. Attendees will learn that working with large data sets is challenging but can provide valuable information about nursing care costs, quality, performance, effectiveness, and outcomes.
3. Attendees can engage with panelists on practice variation that is supported for clinical and translational research applications.

Introduction of the panelists
The panel members have a long-standing collaboration in clinical research informatics and the development of the Nurse Value methodology. Each has decades of interest and research dedicated to building a framework to better understand how nursing care is expended for each person/patient and measure the variation in clinical practice specific to defined outcomes. When key nursing data is gathered in a standardized, interoperable manner, it makes it possible to use the information to identify patterns and trends within nurses, units, facilities, and populations.

Working Group Affiliation: Endorsement of this panel presentation is provided by the Nursing Informatics Working Group (NIWG).

Acknowledgment: The study used in this panel discussion was supported by grants from the Agency for Healthcare Research and Quality (AHRQ) 1R03 HS025495-01 and the University of Colorado School of Medicine Data to Value (D2V) project and The University of Kansas School of Nursing Faculty grant.

All participants have agreed to take part in the panel.

References
Advancing the Collection and Integration of Patient-reported Outcome Data Implementation Architectures Using FHIR® Technical Specifications

Chun-Ju Hsiao, PhD1, Stephanie Garcia, MPH2, Daniella Meeker, PhD3, Joseph Blumenthal, BA4, Keith Marsolo, PhD5

1Agency for Healthcare Research and Quality, Rockville, MD; 2Office of the National Coordinator for Health Information Technology, Washington, DC; 3University of Southern California, Los Angeles, CA; 4MedStar Health National Center for Human Factors in Healthcare, Washington, DC; 5Duke University, Durham, NC

Abstract

The electronic collection of patient-reported outcomes (PROs) supports patient-centered outcomes research by offering a complementary perspective to clinician assessments, and may provide greater insights into health status, symptom burden, adherence, and quality of life. However, PRO data are not routinely available electronically for clinical care or research due to various challenges including the lack of standards and easy-to-use data collection applications. This session will present findings from a project jointly led by the Agency for Healthcare Research and Quality and the Office of the National Coordinator for Health Information Technology that tested the use of the Health Level 7 International (HL7®) Fast Healthcare Interoperability Resources (FHIR®) technical specifications to collect and integrate standardized PRO physical function data in ambulatory care settings. The panel will engage in a discussion regarding the advantages and disadvantages of the different implementation approaches and architectures that were used in the project’s pilot tests as well as opportunities for the scale and spread of these implementation models to other use cases. The learning objectives include: (1) understanding different technical approaches for PRO implementation using FHIR® in settings with different characteristics and resources, and (2) learning about the required capabilities and resources for standardized PRO implementation.

Panel Description

The patient’s perspective is central to healthcare decisions affecting prevention, diagnosis, treatment, and long-term care. As the healthcare industry moves to provide tailored healthcare through initiatives such as precision medicine, patient-centered outcomes research (PCOR) becomes a critical piece of the larger ecosystem, producing new scientific evidence that can help patients, families, and their healthcare providers make informed healthcare decisions. The effective use of patient-reported outcomes (PROs) can play a critical role in improving health care delivery and patient experience with care1-3. Use of PROs for guiding and improving care is especially important for patients with multiple chronic conditions, older adults, and people with disabilities as PROs can reflect outcomes of care for multiple conditions, treated by multiple providers, across multiple settings of care.

While various efforts promote the use of PROs4-5, the uptake is far from universal. Some electronic health record (EHR) systems are able to capture some structured PRO data, but this information is not typically collected at the point of care. Data element and data capture standards could allow for PRO assessments to be conducted and easily shared across clinical or research system, regardless of the EHR or health IT solution being used.

The Agency for Healthcare Research and Quality (AHRQ) partnered with the Office of the National Coordinator for Health Information Technology (ONC) on a project that aims to advance the collection and integration of standardized PRO data that would support interoperable sharing of this information. This joint project was funded through the Patient-Centered Outcomes Research Trust Fund, administered by the Assistant Secretary for Planning and Evaluation of the U.S. Department of Health and Human Services (HHS). In this project, ONC developed a Health Level 7 International (HL7®) Fast Healthcare Interoperability Resources (FHIR®) PRO implementation guide (IG)6 that guides developers in implementing FHIR® for the exchange of PROs as well as the capabilities that are required for a successful integration of PROs into the patient record. Subsequently, AHRQ worked with MedStar Health and ONC worked with two research networks, patient-centered SCAlable National Network for Effectiveness Research (pSCANNER), and Research Action for Health Network (REACHnet), to pilot test applications that use the FHIR® technical specification to collect and send PRO physical function data in ambulatory care settings.
This panel will feature presentations by AHRQ and ONC programmatic leads, and representatives from two pilot test sites. Each pilot site leveraged the NIH-funded Assessment Center application programming interface (API) technology to collect PROMIS® physical function data. The questionnaires and the responses were mapped into a FHIR® format and stored within a FHIR® server. Each pilot site employed a different approach to facilitate PRO data integration with the organization’s respective EHR systems to make the PRO data available to providers. The presentation of pilot tests will be followed by a discussant-led conversation, intended to stimulate thoughts about opportunities in using the FHIR® PRO implementation guide to facilitate the collection and integration of standardized PRO data, with a particular focus on how best to work with different practices to achieve successful implementations and use these data for clinical care, PCOR, and research broadly.

Chun-Ju Hsiao, PhD (moderator and organizer) is a health services researcher within the Health IT Division at AHRQ. She has published research on the topics of electronic health record systems, quality of care, quality of life, disability, health services utilization in office-based physician practices and by nursing home residents, and survey methodology. Dr. Hsiao will introduce the speakers and the topic including the rationale and motivation for bringing this panel together. She will also provide an overview of this joint project and AHRQ’s related tasks including a challenge competition and MedStar Health’s contract work.

Stephanie Garcia, MPH (panelist and co-organizer) oversees the ONC PCOR portfolio, which is a suite of projects that invests in the health information technology infrastructure that supports the health research enterprise. She has collaborated with public and private stakeholders, successfully executing projects that advance the effective use of health information technology to improve the nation's health. Ms. Garcia’s remarks will focus on the development of the FHIR® PRO IG, the required capabilities for PRO integration, and how the IG can be tailored for other measures which may include social determinants of health. She will also provide an overview of ONC’s IG pilot-testing approaches.

Daniella Meeker, PhD (panelist) is an Assistant Professor in the Department of Preventive Medicine at University of Southern California (USC) Keck School of Medicine. She has contributed to standards development for research networks and PROs. She is currently the Informatics Program Director for the Southern California Clinical Translational Sciences Institute Informatics Program, a co-principal investigator of the pSCANNER clinical data research network, and the faculty sponsor of the Los Angeles County Department of Health Services Informatics Core Facility. Dr. Meeker will discuss the challenges and lessons learned of using the FHIR® PRO IG and incorporating PROs into an EHR system. pSCANNER’s solution includes leveraging Clinical Decision Support (CDS) Hooks and integrated services with a SMART on FHIR® app so that providers can order PRO measures.

Joseph Blumenthal, BA (panelist) is a clinical informatics developer with the MedStar Health National Center for Human Factors in Healthcare. Mr. Blumenthal works on process improvement projects to improve patient safety both at MedStar Health and across the country. Mr. Blumenthal will present a novel framework for a system-agnostic implementation of PRO data collection which leverages SMART on FHIR® technology in support of a patient-centered workflow The presentation will highlight creative solutions to the often resource-intensive and protracted process of integration into different EHRs. Mr. Blumenthal will also present the design of a provider-facing visualization to present PRO data in the EHR.

Keith Marsolo, PhD (discussant) is faculty in the Department of Population Health Sciences in the Duke University School of Medicine. His research falls in to two major areas: infrastructure to support the use of EHRs and other real-world data sources in observational and comparative effectiveness research, and standards and architectures for multi-center learning health systems. He participates in the Patient-Centered Outcomes Research Institute’s (PCORI) National Patient-Centered Clinical Research Network (PCORNet), serving as a co-investigator in the PCORNet Coordinating Center. He led efforts to develop version 4.0 of the PCORNet Common Data Model (CDM), which overhauled the way PRO data are stored in the CDM. Dr. Marsolo was a presenter in AMIA 2018 Informatics Summit panel “Capacity Building for PRO Data Sharing across the Federal Networks: EHR Data Quality, Modeling, and Harmonization.” Dr. Marsolo will lead the panel and the audience, in a conversation about challenges and opportunities in using FHIR® to collect and integrate PRO data, focusing on considerations of different IT approaches and architectures. Discussion questions may include:

- What are the pros and cons of different technical approaches and architectures presented today?
- What was required in terms of expertise and resources to implement standardized PRO data collection and integration?
• How might you use the FHIR® PRO IG for domains other than physical function?
• What challenges may occur when utilizing the FHIR® PRO IG?
• What are the pros and cons of different technical approaches and architectures presented today?
• How could you decide which model works better for your organization?
• What considerations your organization will need to take into account in order to adopt any of the models presented today? What kind of technical assistance may be needed?
• What are some potential impacts on workflow?

Topic Rationale:
Increasing value in healthcare by empowering patients with data and information to help them take an active role in making decisions about their care is an HHS priority. The patient perspective is central to health care decisions affecting prevention, diagnosis, treatment and long-term care. PROs critically inform patient-centered outcomes research and can inform clinical management of individuals, shared decision making, patient self-management support, care planning, goal setting, and goal attainment. There is a clear need to understand implementation challenges and find methods to mitigate them. This panel features approaches that can make PRO data available under different health IT infrastructures. Patients, providers, and researchers can use these data for clinical care and research purposes.

Statement of Agreement to Participate
All panelists have approved of this submission and agreed to participate.

References
Educational Communities in the Academic Forum: Sharing Knowledge and Promoting Standards

Stephen Johnson, PhD, FACMI, New York University Langone Health, New York, NY

Saif Khairat, MPH, PhD, School of Nursing, University of North Carolina, Chapel Hill, NC

Suzanne Boren, MHA, PhD, Department of Health Management and Informatics, University of Missouri, Columbia, MO

Vishnu Mohan, MD, MBI, FAMIA Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR

E. LaVerne Manos, DNP, RN-BC, FAMIA, Center for Health Informatics, School of Nursing, University of Kansas, Kansas City, KS

William Hersh, MD, FACMI, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR

Abstract

The mission of the AMIA Academic Forum is to foster the development, growth and improvement of academic units that educate students and conduct research in the discipline of biomedical and health informatics. The Forum provides a vehicle for academic units to interact, consider, and act jointly on issues important to them. This didactic panel describes the new organizational structure of the AMIA Academic Forum based on “educational communities”. The learning objectives for the panel include: 1. Understand how educational communities enable Forum members to engage with their peers to share knowledge, collect data, and promote standards across their institutions. 2. Understand the goals and activities of five different communities for informatics education: baccalaureate, professional master’s, postdoctoral clinical fellows, nursing, and biomedical informatics academic leaders. 3. Understand the benefits for creating and participating in new educational communities within the Forum.

Description

The moderator (Johnson) is the current chair of the Academic Forum, and will briefly describe the goals of the Forum and the recent organizational change to develop educational communities. [1] Mohan will describe the current state of the Clinical Informatics Program Directors community, and discuss recent activities and future goals. [2] Manos will describe the Nursing Informatics Program Directors community, and discuss past achievements as well as new work related to accreditation. [3] Khairat will describe the Baccalaureate Education Community, which is currently being formed, and discuss potential goals and activities. Boren will describe the Professional Master’s Community, which is also being formed, and discuss the relationship to related AMIA efforts regarding accreditation. [4] Hersh will discuss the goals and activities of the community for the Academic Leaders Community.

The panel sessions will consist of five 7-minute presentations, each followed by five minutes of questions. The last 30 minutes of the panel will be devoted to closing discussion, using a list of discussion points to increase audience participation.
Rationale

The topic of this panel is important to update and disseminate to AMIA members the recent organizational changes to the Academic Forum including the expansion into educational communities of practice and benefits of Academic Forum membership. The anticipated audience of AMIA members, academic leaders, program directors and educators can benefit from learning about educational communities to enhance networking with their peers.

Discussion Questions

1. What should be the role of the Academic Forum going forward?
2. How can the AF help programs and faculty members? Examples include network, educational pipeline, teaching collaborations, and building resources or capacity.
3. How can educational communities best serve the needs of individual and organizational members?
4. Which community or communities are most relevant to your institution or program?
5. Which educational communities should the Academic Forum focus on creating next?
6. What should be the relationship between the Academic Forum and the Education Working Group?
7. What should be the relationship between the Academic Forum and other Working Groups?
8. How could industry become more engaged in the Academic Forum?

Agreement to Participate

All participants have agreed to take part on the panel.

Rethinking Health Data Privacy

Bonnie Kaplan, PhD, FACMI¹; Elizabeth J. Davidson, PhD²; George Demiris, PhD, FACMI³; Richard Schreiber, MD, FACP, ABPM-CI⁴; Ari Ezra Waldman, PhD, JD⁵
¹Yale University, New Haven, CT; ²University of Hawaii at Manoa, Honolulu, HI; ³University of Pennsylvania, Philadelphia, PA; ⁴Geisinger Holy Spirit, Camp Hill, PA; ⁵New York Law School, New York, NY

Abstract

Privacy is protected both ethically and legally as a foundation for creating enough trust during clinical encounters for health care decisions to be based on honest discussion and accurate information exchange. Research, public health surveillance, advances in artificial intelligence (e.g., machine learning and predictive algorithms for health care), and the promise of personalized medicine all depend on accurate, complete data. However, health-related data generated and used outside of clinical settings is not protected through privacy regulation. Data aggregators and companies combine data from multiple sources for health and other purposes, while individual behavioral and social practices are being incorporated into medical records. Meanwhile, boundaries and distinctions are breaking down between different categories of protected health data, and between protected data and data collected via commercial apps and services. Privacy of health-related data requires rethinking in this rapidly changing landscape. Panelists will consider challenges in health data privacy, data governance, and privacy policies and practices, including privacy vs the value of data sharing, the adequacy of current legal and regulatory regimes, and how technological developments affect health data privacy.

WG Endorsements: ELSI, POI, CIS

Panel Description

Privacy is a linchpin of the doctor-patient relationship and a foundation for creating enough trust during clinical encounters for health care decisions to be based on honest discussion and accurate information exchange. Research, public health surveillance, advances in artificial intelligence (such as machine learning and predictive algorithms for health care), and the promise of personalized medicine all depend on accurate, complete data. Without privacy, data, as well as care, is compromised.

Health data is protected in the US both by contracts and by law. For example, commercial devices such as wearables, mHealth applications, on-line genetic services, and insurance, are governed by companies’ privacy policies and click-through agreements, enforceable by the Federal Trade Commission, which also enforces the Children's Online Privacy Protection Act (COPPA). A variety of sectoral laws, most notably HIPAA for medical data, the Common Rule for research data, and GINA for genetic data, protect medical data collected primarily in clinical settings. Under these laws, different sources and kinds of data are governed differently, even if covered by the same law (e.g., potentially stigmatizing data related to mental health, substance abuse, and sexually transmitted disease are more protected than other clinical data, while the military and VA have their own practices). Both US and EU data protection are based on Fair Information Practice Principles (FIPPs) that emphasize privacy as an individual’s ability to control data about that person. This depends on the person’s knowing what data is collected, by what entity, for what purposes, for how long, and how it is secured. These are addressed through notice and consent policies, permissions for access, and the ability to correct and remove data. Is the current regimen sufficient?

The current legal regimen assumes competent, autonomous individuals who are responsible for making appropriate decisions, and organizations responsible for compliance through policies that check the boxes of the FIPPs and thereby treat privacy requirements as a ceiling rather than floor for data protection. Further, laws and regulations rely on de-identification or anonymization, yet as data from multiple sources gets combined, it becomes easier to re-identify individuals. Being able to share patient records across settings benefits patients and their health care providers, but depends on identification and thereby makes the information more vulnerable. Identification makes it easier to link data about one individual for such valuable purposes as research and personalized medicine, but also makes privacy harder to protect.

Meanwhile, boundaries and distinctions are breaking down between the different categories of protected data, and between protected data and data collected via commercial apps and services. When wearable device data becomes part of an EHR, or EHR data is mined for business purposes as well as research purposes, or shared across a variety of patient portals and social media, how should data and data uses be governed? How well is privacy protected when medical and pharmacy records are sold and aggregated with data from multiple sources for purposes as disparate as selling advertisements, assessing credit-worthiness, determining insurance eligibility and rates, evaluating potential employees, personalizing diet
advice, policing and crime prevention, pharmaceutical marketing, public health surveillance, and further secondary and tertiary uses, even if originally de-identified?

Moreover, not all data related to health care is legally protected. Examples include: Facebook’s review of posts to flag potential suicides and detect flu outbreaks; Internet of Things applications that issue alerts when assisted living residents may need help; and patient portals that encourage individuals to upload data from on-line and mobile phone apps and medical devices. Especially with the growing recognition of the importance of behavioral and social practices, the promise of personalized medicine, and the incorporation of related data into medical records and healthcare practices, more and more data is potentially health data. However, data often is decontextualized when moved from one place to another, making it harder to assess social, behavioral, and health practices.

The law can do only so much and has difficulty keeping up with rapidly changing social and technological norms. Currently, both privacy and beneficial uses of health data are threatened. This session will rethink privacy law and also privacy policies and practices. Participants and audience will discuss crucial questions, such as:

1. What can law and regulation achieve, what are limitations of regulation given the market value of health data and new technological developments, and what more is needed to ensure privacy?
2. Is sectoral data protection sufficient in light of the combining of all kinds of data sources and means of collection for multiple purposes that may include health?
3. How protective is the notice-and-consent approach? Can privacy policies be protective? How well do they fulfill the goals of notice or consent? Must they be, as commonly is the case, legalistic and non-negotiable (the user either accepts the policy or doesn’t become a user) to the point where virtually no one reads or understands them, if they exist at all? What alternative approaches are viable?
4. Is it effective to protect patient-generated data as part of medical records or tethered patient portals when that same data is shared across social media or collected and sold by private entities?
5. How protective are technical approaches such as de-identification, encryption, and granular permissions? How might privacy by design be implemented? What new approaches are being developed?
6. How well does de-identification serve individual privacy or research and health care needs? How can research, public health, and personalized medicine be better served while protecting privacy?
7. How can health data maintain value when sharing it between clinical encounters and other settings strips it of clinical context, or if collected as part of daily activities, remove significant aspects of that individual’s life from understanding of the data? How does decontextualization as data is removed from its original setting, means, and sources of collection, affect its interpretation, accuracy, and inference made from it? How will privacy norms that change with context be preserved when context changes?
8. What protections should individuals who do not fit the ideal of autonomous, rational actors have? Are special protections needed for the elderly, the mentally ill, people who are very ill, and others more vulnerable? How should teenagers’ privacy be protected?
9. How can health care delivery organizations, businesses, and the general public incorporate practices more in line with public expectations for privacy and the role of privacy in health care?
10. How should the need for privacy be reconciled with the value of data for research, public health, learning health systems, and innovation? How will artificial intelligence advances that could produce new health insights affect health data governance and challenge the current regime? To what degree should the individual’s right to privacy be primary, when sharing data has societal benefits?
11. What non-legal practices and actions (e.g., education, privacy design, social practices) can improve privacy while facilitating beneficial data sharing?
12. How can privacy be reconceptualized to provide a better approach to data protection and to health care? What do new theories, laws, and practices offer?

Discussion will move from front-line experience to empirical work, theoretical models, and potential solutions. After Dr. Kaplan introduces the panel and issues to be discussed, Dr. Schreiber will provide an overview of an on-the-ground discussion of how health care organizations such as the Geisinger Health System address privacy and data governance dilemmas concerning patient portals, OpenNotes, Geisinger’s work with Apple, and teenaged patients. Dr. Demiris then will consider how evidence-based needs of older adults and persons with varying degrees of cognitive ability can expand the concept of “privacy” to concerns such as intrusiveness of technology, burden on users and families, and patient safety. Next, Dr. Davidson will draw on legal and organizational theories and her empirical studies to address how different stakeholders’ views of privacy and rights can be protected or threatened when privacy norms and data governance change as data is shared across settings and used for a variety of purposes, including artificial intelligence and machine learning. Lastly, based on his ethnographic and legal studies, Dr. Waldman will propose a rethinking of health data privacy based on trust and transparency to address problems with current laws, regulations, policies, and their implementation; disparity between views of privacy by privacy officers, technologists and programmers, and the public; how routinized privacy practices
subvert goals of privacy regulation and policy; and how these issues affect health care data. Finally, Dr. Kaplan will summarize the presentations and suggest the need to address privacy for health-related data from multiple perspectives.

Participants

Elizabeth J. Davidson, PhD, professor in the Department of Information Technology Management, Shidler College of Business, University of Hawaii Manoa, serves as Editor-in-Chief of Information and Organization. Her publications include studies of the adoption and diversity of use of integrated clinical systems into hospitals and small physician practices, the evolution of personal health records into patient portals and mobile health apps, and AI and personal health data governance models and cases. She and collaborators are studying the implications of wearable health monitors on physician-patient communication; the integration of telehealth, mobile-enabled glucose monitors, and text messaging in chronic care management in low SES rural communities; the integration of smart technologies in family care for persons with dementia; and health data governance organizations, focusing on the evolution and development of all payer claims databases (APCDs).

George Demiris, PhD, FACMI, Penn Integrates Knowledge University Professor in the School of Nursing and Department of Biostatistics, Epidemiology and Informatics in the Perelman School of Medicine, University of Pennsylvania. In addition to his many papers, he is co-author of Behavioral Intervention Research in Hospice and Palliative Care: Building an Evidence Base (Academic Press, 2018). A Fellow of the Gerontological Society of America, his expertise includes designing and evaluating “smart home” solutions for aging, and understanding the potential of wearable devices or digitally augmented residential settings to facilitate passive monitoring and support independence and quality of life for community dwelling older adults. His research provides evidence-based recommendations for designing systems that are easily adopted by older adults and integrated in their lives.

Bonnie Kaplan, PhD, FACMI, (panel organizer, moderator), of the Yale Center for Medical Informatics, is a Yale Interdisciplinary Bioethics Center Scholar, a Faculty Affiliated Fellow of the Yale Law School's Information Society Project, and Faculty Affiliate of Yale’s Solomon Center for Health Law and Policy, Yale University. Her research addresses informatics ethical, legal, and privacy issues; user perspectives and experiences with health information technology; and ethnographic sociotechnical evaluation. Among her many publications and presentations, are journal papers and book chapters on health data privacy and on socio-technical approaches for cybersecurity and for evaluating health information systems. She twice served as chair of the AMIA Ethical, Legal, and Social Issues Working Group; AMIA's People and Organizational Issues Working Group; and the IMIA Organizational and Social Issues Working Group.

Richard Schreiber, MD, FACP, ABPM-CI, is Associate Chief Medical Informatics Officer, Geisinger Health System, and Chief Medical Informatics Officer at Geisinger Holy Spirit (GHS) in Camp Hill, PA, and Regional Assistant Dean of the Geisinger Commonwealth School of Medicine. He is board certified in Clinical Informatics and Internal Medicine, and a Fellow of the American College of Physicians. He spearheaded physician EHR adoption at GHS, and after 10 years of successful use, then led the transformation to a system-wide EHR which is essentially paperless. He is Vice-Chair of AMIA’s Clinical Informatics Systems Working Group. His clinical informatics research focuses on clinical decision support, CPOE, drug-drug interaction alerting, documentation improvement, problem list curation, and CDS anomalies.

Ari Ezra Waldman, PhD, JD, is a Professor of Law and the Director of the Innovation Center for Law and Technology at New York Law School. An internationally recognized thought leader on privacy and online safety, he is the Founder and Director of the Institute for CyberSafety. A sociologist and legal scholar, he studies how law and technology mediate social life, focusing on privacy, technology design, and speech online. His first book, Privacy As Trust: Information Law for an Information Age (Cambridge University Press, 2018), argues that privacy law should protect information disclosed in contexts of trust.
The Path Forward: Recommendations from the U.S. Health IT Advisory Committee on Interoperability Standards Priorities and the U.S. Core Data for Interoperability

Kensaku Kawamoto, MD, PhD, MHS1, Steven Lane, MD, MPH, 2
Christina Caraballo, MBA,3 Terrence O’Malley, MD4

1Department of Biomedical Informatics, University of Utah, Salt Lake City, UT,
2Sutter Health, Sacramento, CA,
3Audacious Inquiry, Baltimore, MD
4Partners HealthCare, Boston, MA

Abstract

While substantial progress has been made in recent years to improve health care through standards-based interoperability, much remains to be done. The U.S. Health IT Advisory Committee (HITAC) was established in the 21st Century Cures Act of 2016 to provide recommendations to the National Coordinator for Health IT to advance the electronic access, exchange, and use of health information. In this session, the co-chairs of the HITAC U.S. Core Data for Interoperability (USCDI) and Interoperability Standards Priorities (ISP) Task Forces will describe the HITAC recommendations for advancing interoperability in the U.S. in these areas. These recommendations encompass the process for expanding the data elements included in the USCDI; the specific data elements with highest priority for near-term inclusion in the USCDI; the priority uses of health IT; gaps in standards or their implementation that hinder these priority uses; and recommended industry and government action to address these gaps. Following an overview of the HITAC recommendations in these areas, the panelists will engage in an interactive dialogue with the audience on what is most needed to advance health care in the U.S. through standards-based interoperability.

Description

The overall learning objective of the panel is for the learner to be able to describe key needs and potential solutions for standards-based interoperability in the U.S., in particular with regard to the interoperability needs identified by the HITAC’s U.S. Core Data for Interoperability (USCDI) and Interoperability Standards Priorities (ISP) Task Forces. The panelists will pursue achievement of this learning objective by describing the recommendations of the HITAC in these areas for improving health care in the U.S. through standards-based interoperability.

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>Kawamoto</td>
<td>Introduction to the panelists</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Need for improved standards-based interoperability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overview of U.S. Health IT Advisory Committee (HITAC)</td>
</tr>
<tr>
<td>15 min</td>
<td>Caraballo</td>
<td>Charge and membership of the U.S. Core Data for Interoperability (USCDI) Task Force</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HITAC recommendations on USCDI (Part 1)</td>
</tr>
<tr>
<td>15 min</td>
<td>O’Malley</td>
<td>HITAC recommendations on USCDI (Part 2)</td>
</tr>
<tr>
<td>15 min</td>
<td>Kawamoto</td>
<td>Charge and membership of the Interoperability Standards Priorities (ISP) Task Force</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HITAC recommendations on ISP (Part 1)</td>
</tr>
<tr>
<td>15 min</td>
<td>Lane</td>
<td>HITAC recommendations on ISP (Part 2)</td>
</tr>
<tr>
<td>25 min</td>
<td>All</td>
<td>Panel discussion with audience</td>
</tr>
</tbody>
</table>
**Kawamoto:** Dr. Kawamoto is Associate Chief Medical Information Officer of University of Utah Health, as well as Associate Professor and Vice Chair for Clinical Informatics in the University of Utah Department of Biomedical Informatics. Dr. Kawamoto is also co-chair of the HL7 Clinical Decision Support Work Group, a Board Member of Health Level 7 International (HL7), and co-chair of the HITAC ISP Task Force. Dr. Kawamoto also leads the University of Utah’s ReImagine EHR initiative, which is a multi-stakeholder effort to improve patient care and the provider experience through standards-based, interoperable extensions to the EHR.

Dr. Kawamoto will serve as the moderator and introduce each of the panel members and their role in the HITAC. Dr. Kawamoto will then briefly describe the central issue at hand: the need for improved standards-based interoperability to support more efficient and effective health care. He will then provide an overview of the HITAC, which was established in the 21st Century Cures Act of 2016 to provide recommendations to the National Coordinator for Health IT to advance standards-based interoperability. The HITAC unifies the roles of, and replaces, the Health IT Policy and Standards Committees that existed prior to the enactment of the 21st Century Cures Act. HITAC members are appointed by the U.S. Government Accountability Office as well as by the Secretary of the U.S. Department of Health and Human Services, the majority and minority leaders of the U.S. Senate, and the Speaker and minority leader of the House of Representatives. As of March 2019, the HITAC consists of 26 members who represent stakeholders including patients, healthcare providers, and technology vendors. The HITAC conducts much of its work through Task Forces. Each task force is co-chaired by HITAC members and is comprised of both HITAC members and non-HITAC subject matter experts. Task Forces present draft recommendations to the full HITAC, which are then discussed, revised as needed, and then adopted as formal recommendations of the HITAC. HITAC recommendations are provided to the Office of the National Coordinator for Health IT (ONC), which considers the recommendations in making policies and furthering standards-based interoperability. Specific policies for which HITAC recommendations are considered include proposed rules on Interoperability, Information Blocking, and the ONC Health IT Certification Program pursuant to the 21st Century Cures Act.

**Caraballo:** Ms. Caraballo is Director of Audacious Inquiry and co-chair of the HITAC U.S. Core Data for Interoperability (USCDI) Task Force. Ms. Caraballo is a leading global subject matter expert on consumer engagement strategies and initiatives and has been at the forefront of the global culture shift to put patients and families at the center of care. Ms. Caraballo acts as an advisor to the Department of Commerce on global exports of health IT, is on the Board of Directors for the National Association for Trusted Exchange, is the Chair of the HIMSS Technical Guidance for Interoperability Work Group, and was one of seven inaugural recipients of the HIMSS 2016 Most Influential Women in Health IT Award.

Ms. Caraballo will provide an overview of the charges of the USCDI Task Force, which broadly encompass formulating recommendations on how the USCDI is to be expanded, as well as on what should be included in the USCDI in its next stages. The USCDI is a standardized set of health data classes and constituent data elements for nationwide, interoperable health information exchange. Following the overview of the Task Force charge, Ms. Caraballo will briefly describe the membership of the Task Force, which includes HITAC members as well as external subject matter experts. Ms. Caraballo will then describe the recommendations that were made to the HITAC, and subsequently adopted following some modification, for the Promotion Model Lifecycle of proposed data elements. This lifecycle includes articulation of the need for the data element, the weighing of benefits and costs, the development and refinement of standards, and real-world implementation and validation of the proposed standards.

**O’Malley:** Dr. O’Malley is a geriatrician at Massachusetts General Hospital with experience providing care to patients with complex medical needs. Together with Ms. Caraballo, he serves as co-chair of the HITAC USCDI Task Force. Dr. O’Malley is an instructor at Harvard Medical School, and he previously held various administrative leadership positions, including as the Medical Director for Non-acute Care Services at Partners HealthCare and Partners HealthCare at Home. His areas of expertise are in quality measurement and process improvement, systems design, and clinical care, particularly in long-term and post-acute settings.

Dr. O’Malley will describe the HITAC recommendations on the data elements that should be included in the USCDI in the next phases. These deliberations are currently ongoing, with a specific focus on data elements including clinical notes, provenance, and pediatric data elements. Dr. O’Malley will describe the recommendations that result from these deliberations.

**Kawamoto:** Dr. Kawamoto will provide an overview of the charge of the ISP Task Force, which is to formulate recommendations on priority uses of health IT and the associated standards and implementation specifications that
support such uses. He will also briefly describe the membership of the Task Force, which includes representatives from Epic, Cerner, IBM, Apple, and Google. Dr. Kawamoto will then describe the process used by the Task Force to identify and rank the priority use cases, as well as the results of that prioritization process: 1) Orders and Observations; 2) Referrals and Care Coordination; 3) Medication and Pharmacy; 4) Point-of-Care Guidance (including evidence-based disease management, prior authorization, and price transparency); and 5) Social Determinants of Health. Dr. Kawamoto will then describe the process by which the ISP Task Force identified standards gaps in these prior use cases and formulated recommendations with the assistance of invited subject matter experts. He will outline draft ISP Task Force recommendations in the area of point-of-care guidance.

**Lane:** Dr. Lane is Clinical Informatics Director for Privacy, Information Security & Interoperability at Sutter Health, and he serves as co-chair of the HITAC Interoperability Standards Priorities (ISP) Task Force. Dr. Lane is also a practicing primary care physician and has been a leader in advocating for interoperability in California and around the country through lowering barriers to information sharing and optimizing the utility of the data exchanged. Dr. Lane has served as a member of various state and national-level work groups focused on interoperability, privacy, and information security, including the Certification Commission for Healthcare Information Technology (CCHIT). He currently serves on the boards of Carequality and The Sequoia Project, as a member of the HL7 Da Vinci Project Steering Committee, the international Epic Care Everywhere Governing Council, co-chair of the DirectTrust Clinicians Steering Workgroup, and Clinical Professor of Family & Community Medicine at UCSF.

Dr. Lane will describe the adopted and draft recommendations of the HITAC in ISP priority areas other than point-of-care guidance. These recommendations will encompass those areas for which recommendations have already been completed (priority areas 1 and 2 above), as well as those areas for which recommendations are under development (the remaining priority areas, except for the point-of-care guidance topic covered by Dr. Kawamoto).

Following panelist presentations, Dr. Kawamoto will lead a moderated discussion with the audience. The objectives of this discussion will be to answer questions from the audience, engage in a stimulating exchange of ideas, and receive feedback on the recommendations of the HITAC Task Forces co-chaired by the panelists. The questions listed below will be presented to stimulate this discussion, and the audience will be encouraged to engage in discussion in areas that are most important to them.

**Significance of panel topic and anticipated audience**

With the rapid adoption of EHR systems in recent years, there is significant potential for electronic data and health IT systems to improve patient care and reduce clinician burdens. Yet, due in no small part to challenges with standards-based interoperability, the gap between the promise and reality of health IT remains significant. Thus, it is imperative for biomedical informaticists to understand the gaps that still exist regarding standards-based interoperability, as well as national-level efforts to address these gaps.

**Discussion questions**

How do you think data elements should be added to the U.S. Core Data for Interoperability?

Which data elements do you think should be added to the U.S. Core Data for Interoperability?

What do you think are the priority uses of health IT? In particular, what are priority uses related to health information exchange?

What recommendations do you have for addressing gaps in standards or their implementations for these priority uses?

What challenges have you encountered in adopting interoperability standards?

**Participation statement**

All proposed panelists have agreed to participate in the panel if the proposal is accepted.
Lessons Learned Using Standards-Based Decision Support Frameworks in Commercial EHRs: Opioid Decision Support Case Study

Kensaku Kawamoto, MD, PhD, MHS¹, Bryn Rhodes², Nitu Kashyap, MD³, Cole Erdmann⁴

¹Department of Biomedical Informatics, University of Utah, Salt Lake City, UT,
²Dynamic Content Group, Orem, UT,
³Yale New Haven Health, New Haven, CT,
⁴Cerner, Kansas City, MO

Abstract

Commercial electronic health record (EHR) systems are beginning to provide support for standards-based clinical decision support (CDS) frameworks in their general-release software. Initial support is focusing on the HL7 Clinical Decision Support (CDS) Hooks standard, and some EHR vendors are beginning to support the HL7 Clinical Quality Language (CQL) standard as well. EHR vendors are also supporting the US Core Fast Healthcare Interoperability Resources (FHIR) profiles for data access. This panel will describe lessons learned from a Centers for Disease Control and Prevention (CDC) and Office of the National Coordinator for Health IT (ONC)-sponsored effort to pilot the use of opioid CDS knowledge resources using these EHR-supported CDS interoperability frameworks. Important lessons learned include the need for EHR support for additional CDS Hooks trigger events, in particular for medication ordering; the still early nature of direct CQL support in EHR vendor systems; the need to combine standards-based CDS with supplemental logic implemented using EHR platform-specific CDS rules engines; and the need for expansion and EHR vendor support for US Core FHIR profiles and interfaces, including for placing orders and for retrieving information on dispensed medications.

Description

The overall learning objective of the panel is for the learner to be able to understand the current state of standards-based CDS frameworks as well as their degree of support in commercial EHR systems. The panelists will pursue achievement of this learning objective by describing their experiences attempting to leverage these EHR vendor-supported frameworks to address the opioid epidemic.

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 min</td>
<td>Kawamoto</td>
<td>Need for interoperable CDS&lt;br&gt;Overview of HL7 CDS Hooks, CQL, and US Core FHIR standards&lt;br&gt;Initiative to develop and disseminate standards-based CDS for opioid use</td>
</tr>
<tr>
<td>15 min</td>
<td>Rhodes</td>
<td>Details of relevant standards and state of EHR vendor support&lt;br&gt;Description of standards-based knowledge artifacts for chronic pain management with opioids</td>
</tr>
<tr>
<td>15 min</td>
<td>Kashyap</td>
<td>Lessons learned from pilot with Epic EHR and CDS Hooks at Yale University</td>
</tr>
<tr>
<td>15 min</td>
<td>Erdmann</td>
<td>Lessons learned from pilot with Cerner CQL Compiler at Indiana University</td>
</tr>
<tr>
<td>30 min</td>
<td>All</td>
<td>Panel discussion with audience</td>
</tr>
</tbody>
</table>

Kawamoto: Dr. Kawamoto is Associate Chief Medical Information Officer of University of Utah Health, as well as Associate Professor and Vice Chair for Clinical Informatics in the University of Utah Department of Biomedical Informatics. Dr. Kawamoto is also co-chair of the HL7 CDS Work Group, a Board Member of HL7, a member of the U.S. Health IT Advisory Committee (HITAC), and co-chair of the HITAC Interoperability Standards Priorities Task Force. Dr. Kawamoto leads the University of Utah’s ReImagine EHR initiative, which is a multi-stakeholder...
effort to improve patient care and the provider experience through standards-based, interoperable extensions to the EHR.

Dr. Kawamoto will serve as the moderator and introduce each of the panel members and their organizations. Dr. Kawamoto will then describe the central issue at hand, which is the need for interoperable CDS to enable effective CDS at scale. He will then provide an overview of the standards relevant for integrating CDS at the point of care, with a focus on the HL7 CDS Hooks, CQL, and US Core FHIR standards. He will then note that commercial EHR vendors are beginning to support these standards, and that the panel will focus on early, real-life implementation experience with these EHR-supported frameworks to help advance their use and continued maturation. Finally, Dr. Kawamoto will introduce the case study, which is a project sponsored by the Centers for Disease Control and Prevention (CDC) and the Office of the National Coordinator for Health IT (ONC) to facilitate the standards-based implementation of the 2016 CDC Guideline for Prescribing Opioids for Chronic Pain (CDC Guideline).

Rhodes: Bryn Rhodes is Principal of Dynamic Content Group. Mr. Rhodes is also co-chair of the HL7 CDS Work Group, the lead author of the HL7 CQL standard, and a co-author of the HL7 CDS Hooks standard. He is also a core subject matter expert on the CDC-ONC standards-based opioid CDS initiative. Mr. Rhodes will provide details on the relevant standards for which Dr. Kawamoto has provided an overview, using the CDS knowledge artifacts generated for the CDC Guideline, which are CDS Hooks services using CQL as the underlying knowledge representation formalism. The Table outlines the 12 CDS knowledge artifacts that will be developed by September 2019 for this project, with each knowledge artifact corresponding to one of the 12 recommendations in the CDC Guideline. Six of these recommendations have already been completed, and recommendations #10 and #11 are being used for the current pilots.

Table. CDC Recommendations Supported via Standards-Based CDS Knowledge Artifacts

<table>
<thead>
<tr>
<th>#</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selection of non-pharmacologic therapy, nonopioid pharmacologic therapy, opioid therapy</td>
</tr>
<tr>
<td>2</td>
<td>Establishment of treatment goals</td>
</tr>
<tr>
<td>3</td>
<td>Discussion of risks and benefits of therapy with patients</td>
</tr>
<tr>
<td>4*</td>
<td>Selection of immediate-release or extended-release and long-acting opioids</td>
</tr>
<tr>
<td>5*</td>
<td>Dosage considerations</td>
</tr>
<tr>
<td>6</td>
<td>Duration of treatment</td>
</tr>
<tr>
<td>7*</td>
<td>Considerations for follow-up and discontinuation of opioid therapy</td>
</tr>
<tr>
<td>8*</td>
<td>Evaluation of risk factors for opioid-related harms and ways to mitigate patient risk</td>
</tr>
<tr>
<td>9</td>
<td>Review of prescription drug monitoring program (PDMP) data</td>
</tr>
<tr>
<td>10*</td>
<td>Use of urine drug testing</td>
</tr>
<tr>
<td>11*</td>
<td>Considerations for co-prescribing benzodiazepines</td>
</tr>
<tr>
<td>12</td>
<td>Arrangement of treatment for opioid use disorder</td>
</tr>
</tbody>
</table>

*Currently implemented. Others are specified and undergoing implementation (completion = September 2019).

Kashyap: Dr. Kashyap is Associate Chief Medical Information Officer and Executive Director of Medical Informatics at Yale New Haven Health. Dr. Kashyap will describe Yale University’s experiences piloting the standards-based opioid CDS knowledge artifacts using the Epic EHR and its native support for the CDS Hooks “patient-view” hook, which was introduced in its August 2018 software version. This piloting began in February 2019, and includes hosting of a local copy of the CDS Hooks service, ensuring that local data are mapped to the appropriate standard codes, configuring the native CDS rules engine to restrict the CDS Hooks service to only be invoked for appropriate providers and for appropriate patients, and creating and evaluating test patients in the Epic testing environment. Dr. Kashyap will describe important lessons learned, with a particular focus on issues that need to be considered for widespread dissemination. Already, lessons learned have included the need for EHR support for additional CDS Hooks trigger events, in particular for medication ordering; the need to combine standards-based CDS with supplemental logic implemented using EHR platform-specific CDS rules engines; and the need for expansion and EHR vendor support for US Core FHIR profiles and interfaces, including for placing orders and for retrieving information on dispensed medications.
**Erdmann:** Finally, Mr. Erdmann will describe Cerner’s experiences ingesting the opioid CDS knowledge artifacts into the Cerner HealtheIntent platform and the integration with Cerner Millennium running at Indiana University Health. This piloting began in May 2019, and includes direct translation of the CQL logic from the opioid decision support artifacts into an Apache Spark data pipeline, evaluating the decision support for the relevant population to create appropriate recommendations, and displaying those recommendations to providers using Cerner EHR functionality. Lessons learned have included feedback on the way the shareable decision support logic is expressed to support more streamlined ingestion into the Spark environment.

Following these panelist presentations, Dr. Kawamoto will lead a moderated discussion with the audience. The objectives of this discussion will be to answer questions from the audience, engage in a stimulating exchange of ideas, and identify key steps that the informatics community can take to achieve more effective and comprehensive CDS leveraging standards-based interoperability. The questions listed below will be presented to stimulate this discussion, and the audience will be encouraged to engage in discussion in areas that are most important to them.

**Significance of panel topic and anticipated audience**

No single organization has the resources or expertise required to meet all of the CDS needs of health care on its own. The emergence of native EHR support for CDS Hooks, CQL, and US Core FHIR profiles provides a highly promising avenue for healthcare systems, EHR vendors, government agencies, and various other organizations to come together to improve health and health care. At the same time, our initial experiences indicate that there are important challenges that still must be overcome. Thus, it is imperative for biomedical informaticists to understand not only the promise of standards-based CDS interoperability but also the current state of the field and the challenges that must be overcome to achieve this promise.

**Discussion questions**

What experiences have you had with using standards-based CDS at your organization, and what lessons have you learned?
What is your vision for the future of interoperable CDS?
What challenges do you see to achieving this vision?
What steps can you take to help us overcome these challenges?

**Participation statement**

All proposed panelists have agreed to participate in the panel if the proposal is accepted.
Abstract: Biomedical Informatics (BMI) is the pillar of a functioning Learning Health System (LHS) that enables bringing translational clinical research findings into clinical practice and thereby, improving the health of patients and populations. It is a compelling promise; hence, medical centers invest toward establishing sustainable BMI programs in the form of a Center for BMI or an academic division. The Wake Forest School of Medicine (WFSoM) and the Wake Forest Clinical and Translational Science Institute have been striving to create a LHS, and correspondingly, a BMI program to support the successful implementation of a LHS. Wake Forest Center for Biomedical Informatics (WFBMI) attempts to foster learning from care experiences and develop and/or deliver better and more robust algorithmic models (e.g. Clinical decision support [CDS] systems, etc.) utilizing its integrated data ecosystem. In this panel, starting with a historical background, we will discuss challenges and opportunities of establishing a BMI program to support the operation and necessary changes to the institution. Given the demand for BMI services and the associated resource needs of establishing a program, some organizations may benefit from to-be shared experiences in order to alleviate potential struggles for such a costly endeavor.

Enabling A Learning Health System with Biomedical Informatics

Healthcare data introduces unparalleled opportunity to learn from each patient-care event, notwithstanding, the collection, storage, retrieval, communication, and optimal usage of such data present tremendous challenges. Biomedical Informaticians pioneer tools and algorithms to transform data into best practice knowledge that is rapidly available at the point of care. However, due to the variation of data types and the institutional implementation approaches, these data reside in repositories with differing data formats, models, and semantics. BMI aims to tackle the aforementioned challenges via its tri-partite mission. First is the provision of informatics services to the school of medicine or the health center that support the needs of biomedical research and care delivery not usually met by the IT departments. Second is the independent or collaborative scientific research development to continuously capture, curate, and validate new evidence. The last mission is the training and education of the next generation of informatics professionals and translational researchers who may benefit from basic informatics expertise. In this panel, we will discuss challenges and lessons learned for beginning a BMI program and respective missions in the following order:

1- **Dr. Marty Kohn** will moderate the panel and introduce the program with a historical narrative, setting the foundation for the need of BMI programs and centers: As a veteran clinical informatician, Dr. Kohn has been heavily involved in the development of BMI programs across the country and has been a thought leader to our community through AMIA and other venues. He will provide an evaluation of the BMI programs that have obvious similarities to other specialties becoming a self-sufficient academic entity within medical schools.

2- **Dr. Metin Gurcan** will follow with a brief overview of the WFBMI and proceed with a discussion of WFBMI’s research efforts and the vast knowledge resulting from ongoing studies. Building a data ecosystem to support LHS requires established research efforts for evidence generation that provide benefits as well as unique challenges to a medical center. For instance, medical imaging plays a critical role in clinical decision making starting with detection, diagnosis, prognosis and later with treatment and monitoring of diseases. Traditionally, these decisions (e.g. in radiology, pathology, dermatology) have been carried out by expert physicians who subspecialize in various disease groups. Yet, these experts are subject to perceptual and cognitive pitfalls that negatively affect their decision making which unfortunately lead to poor healthcare outcomes. Researchers in WFBMI have been successfully designing, developing, validating, and deploying machine learning algorithms in imaging to help physicians make more accurate and consistent decisions. Importantly, these algorithms produce data that can be readily combined with other sources of data (e.g. -omics, electronic healthcare) to improve the quality of decisions. We will discuss two pilot projects where machine learning is being employed. First is the AutoScope for Acute infections of the middle ear (acute otitis media [AOM]), which is the most commonly treated childhood disease. We have been developing a simple and objective
method of analyzing an image of a patient's ear to diagnose or rule out AOM with the aim to drastically reduce overtreatment. The second project is the Follicular Lymphoma Grading System (FLAGS). Currently, the inter-reader agreement between pathologists in grading Follicular lymphoma is extremely low: in a multi-site study, the agreement among experts for the various grades of follicular lymphoma varied between 61% and 73%. We have developed an effective computer-aided system to assist pathologists in the histological grading of follicular lymphoma. For both AutoScope, and FLAGS, we will discuss the importance of good data hygiene and best evaluation practices while paying attention to design, development, validation and project research team composition.

3- Dr. Umit Topaloglu intends to deliberate upon Operational Informatics and WFBMI services towards establishing an infrastructure that enables repeatable and cost-effective learning. In order to address the computational needs of the researchers and business operations at our institution, WFBMI has invested prudently in elastic computing infrastructures resulting from a contractual engagement and a Business Associate Agreement with Google Cloud Platform (GCP) and Microsoft Azure. GCP serves as the flexible research computing framework with available tools to address the needs of precision medicine and other large-scale computing endeavors. WFBMI also utilizes Microsoft Azure as a common computing model for several data sharing and federated learning projects with three other medical centers in the region.

While the reuse of clinical data is hampered by the still prevailing disconnect between the data standards used in patient care and the ones used in clinical research, some efforts to bridge this gap are finally reaping their benefits in the form of clinical research and Electronic Health Record (EHR) system integrations. Substitutable Medical Applications and Reusable Technologies (SMART) on Fast Health Interoperability Resources (FHIR) is a web based standard platform that enables developing solutions to enhance interoperability with EHRs and other systems, including clinical research data capture (e.g. REDCap). The REDCap has implemented a SMART on FHIR module to bridge the gap between clinical care and data capture systems, improving operational efficiency of clinical research by reducing duplicate data entry and redundant workflows. WFBMI has obtained necessary security and compliance approvals and has begun the pilot implementation of REDCap and Epic integration.

To facilitate a comprehensive and semantically rich data ecosystem, the WFBMI has built a Translational Data Warehouse (TDW). The TDW serves as a central resource for a gamut of research and quality improvement needs in the institution. Data is mapped to the standards and extracted from primary sources via utilizing Extract, Transform, Load (ETL) and FHIR based processes. WFBMI is committed on developing thoughtful semantic frameworks to minimize data misinterpretation and discovery challenges and maintains a UMLS repository and a local terminology system. Investigators can perform queries on the TDW using i2b2 and upon appropriate IRB approval, they can download automatically generated datasets through a self-service tool. In addition to EHR data, the TDW has been linked to outside data sources, most notably US census data, and the NC State Death Registry. TDW is also our conduit for participation in the federated data networks for obtaining larger sample sizes and realization of the full potential of population research. WFBMI enables WFSoM’s involvement and contributes to regional and national clinical data research networks, including a) Carolina Collaborative, b) STAR (Mid-South) PCORNet Clinical Data Research Network, c) Accruals to Clinical Trials (ACT) Network of NCATS, d) TriNetX, and e) AACR Genomics Evidence Neoplasia Information Exchange (GENIE).

4- Dr. Brian Wells will present the educational and training efforts of the WFBMI towards training physicians and non-clinician researchers. Example initiatives include Clinical Informatics Board Certification Subsidies through which practicing clinicians are able to become board certified in Clinical Informatics (American Board of Preventive Medicine) through the practice pathway. After which, diplomates will be required to complete a formal 2-year Clinical Informatics fellowship. WFBMI provides financial support for review course as well as the cost of the exam, through which we have increased the number of board certified clinical informaticians in the medical center to cope with the dynamic clinical informatics needs.

WFBMI also organizes Short Course on Clinical Informatics annually that provides a general overview of medical informatics for academic researchers from diverse backgrounds. The hands-on sessions include courses such as creating a statistical model for predicting the risk of readmission for adults admitted for pneumonia. Each iteration of the short course has involved approx. 32 hours of in-class time spread out over a 3-6-month period and we incorporate feedback into the subsequent sessions.
Another Subsidy WFBMI established is the Epic – Physician Builder Certification for which WFBMI provides $3,000 for each clinician to travel to Epic for the 2-part Physician Builder program. The certification focuses on front-end workflow optimization or analytics areas, serving as necessary authorization for certified clinicians to gain elevated user rights in WFSoM Epic. With new user roles, certified builders can create EHR optimization tools such as standardized data capture forms, CDS at the point of care, and dynamic logic-based order sets, registry frameworks for population health and analytics capabilities to facilitate EHR data extraction. Most importantly, the coursework enhances the clinicians understanding of the architecture underlying the EHR and a broader concept of system’s capabilities. The hope is that this training will empower the institution to take advantage of the continuously growing capabilities of the system. In particular, the WFBMI is interested in supporting clinicians who will use this new knowledge and these new skills for QI, clinical innovations and research related projects that support the realization of the LHS.

WFBMI spearheads another initiative in the form of Clinical Informatics Affinity Group to gather researchers throughout the institution with research or interest in clinical informatics and seek their contribution for improving LHS with informatics. The bidirectional nature of such learning and education opportunity reaches group members with ongoing announcements about upcoming educational and funding opportunities. A smaller core group of researchers meets monthly to discuss hurdles that researchers may face in their informatics related research, to discuss ongoing institutional initiatives, barriers, and to contrive ways to catalyze informatics research at the institution. The group also sponsors periodic “Lunch and Learn” sessions that highlight ongoing research projects or topic areas of interest.

5. Dr. Ajay Dharod will discuss the Clinical Scholars in Informatics (CSI) and how applied informatics training shapes clinical workflows and decision making. The CSI is a two-year applied educational and mentorship pathway for internal medicine residents to develop skills in applied informatics. At residency completion, CSI residents are prepared to tackle most health-system challenges with additional comprehensive understanding of informatics (e.g. CDS, workflow, change management, IT systems, EHR, database structures, database querying, regulatory issues, Security and Privacy, etc.). The CSI pathway translates a resident derived project (addressing system pain points) into reality. Furthermore, the CSI yields mutual benefits for the resident physicians (learning the mechanisms for completing an applied clinical informatics project end-to-end with academic capital) and for the LHS. CSI enables residents to pursue career opportunities in informatics complementing a career in internal medicine specialties or subspecialties. Dr. Dharod will also talk about the WFBMI’s newly established internship program that acquires and mentors undergraduate and graduate students from surrounding colleges to train next generation informaticians.

**Summary:** WFBMI is pragmatically designed to contribute to Wake Forest Baptist Health and School of Medicine’s evolution as a learning health system whilst also enhancing better health across the local and national community by advancing and promoting informatics research, practice and education. This panel will discuss how we have designed the program, the challenges faced along the way, and the cost-effective solutions we have been applying.

**Audience and Timeliness**
The benefit of LHS is becoming more evident, this session will be of interest to all biomedical informatics professionals and administrators who are in process or planning to establish a BMI to support

**Other notes** All participants have agreed to participate to the panel.

**Acknowledgement:** This effort is partly supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number ULLTR001420 and by the Cancer Center Support Grant from the National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30 CA012197). We also thank all the faculty, staff, and students of the WFBMI for their contribution and diligent work.

- **A list of discussion questions**
  - Advantages of designing a BMI program aiming for LHS success
  - Achieving project sustainability through funding
  - Working with institutional IT and effective IT-BMI collaboration
  - Establishing an effective computational infrastructure to address unique needs of BMI
  - Training residents and approaches to consider the residency programs
Maturity Models for Improving Health Information System and Digital Health Capacity in Low Resource Settings: Sharing and Discussing Implementation Experiences

Authors

M. Kumar, MPH, MSc, MEASURE Evaluation, University of North Carolina at Chapel Hill
X. Santas, BA, Associate Director for Informatics and Information Resources, Centers for Disease Control and Prevention, Atlanta
J. Kariuki, MPH, BSc, Centers for Disease Control and Prevention, Atlanta
P. Mechael, PhD, MHS, Principal and Policy Lead, Health Enabled, Washington, DC
A. Waugaman, MA, BA, Senior Advisor, Digital Health, United States Agency for International Development, Washington, D.C.

Abstract

This didactic panel aims to share global and country-specific experiences in implementing national digital health and health information system (HIS) assessment tools to facilitate the transition from data to knowledge to action—namely the Global Digital Health Index, HIS Stages for Continuous Improvement, and the HIS Interoperability Maturity Model. The panel presentations and discussion are targeted toward policy makers, health managers and health informaticians engaged in design and use of digital health technologies to inform data-based decision making at the policy, program and service delivery levels. Discussion is expected to focus on emerging trends and findings from the assessments and what the global health and HIS community can do collectively to address them.

Introduction

While digital technologies are fueling the growth of digital health data, evidence shows the necessity of adopting an holistic approach to facilitate the transition from data to knowledge to action in the healthcare system in order to accelerate the goal of improved patient and population health outcomes.1 In the context of lower- and middle-income countries (LMICs), the healthcare system operates in a complex and resource-constrained environment2 and delivers health care services at multiple levels, i.e., community, county, region and national levels. A tiered health care system demands contextualization of interventions aimed at facilitating the progression from data to knowledge to action. These different levels of the health care system have diverse, often competing, health information needs and structural inequalities in terms of decision-making authority, access to financial resources and a skilled workforce, and availability of technology infrastructure. HIS management often includes competing challenges such as service delivery vs. mandates from program administration, and whether to focus on reporting data up in the hierarchical health care system or to prioritize local use of data for action to improve health outcomes. The diversity within and across the health care system of LMICs reaffirms the value of contextual adaptation of digital health technologies for transforming the work, the health workforce that delivers health care, the enabling digital health infrastructure, and their relationships with each other and with patients3 to achieve the stage of data to knowledge to action which then contributes to improved health outcomes.4 A key area of unmet need within these strengthening efforts is the lack of standardized frameworks, tools, and metrics for holistically assessing progress and to inform further improvement. National governments and institutions need measurable strategies, that are aligned with national and global health priorities, for investing and continuously improving their HIS, digital health capacity, and interoperability among diverse HIS. To meet this need, global expert groups have developed and implemented maturity models5 to guide in-country discussions and establish standard metrics within and across various institutions to enhance HIS performance and digital health capacity. These tools include the Global Digital Health Index, HIS Stages for Continuous Improvement, and the HIS Interoperability Maturity Model.

Panel presentations will share global and country-specific implementation experiences highlighting how countries are identifying gaps and executing appropriate interventions to ensure the enabling environment for the effective use of digital health and HIS that translates data to action. Presentations will also offer insights about how the tools can be
used together in the same country while working with a homogenous group of stakeholders besides engaging participants on potential next step in the adaptation and application of these maturity models and metrics.

**Audience**

The panel presentations and discussion are targeted toward policy makers, health managers and health informaticians engaged in design and use of digital health technologies to inform data-based decision making at the policy, program and service delivery levels. The discussion aims to elicit feedback about the benefits and challenges associated with the implementation of multiple digital health interventions in resource constrained environment of LMICs.

**Speaker Contributions**

The panel consists of four speakers and a moderator. Xenophon Santas will serve as the panel moderator and explain the purpose and content of the panel discussion. Manish Kumar will share an overview of maturity model-based tools and discuss the coordination in the design and development processes. Patricia Mechael will share experiences and trends from the implementation of the Global Digital Health Index. James Kariuki will present the findings of the HIS Stages of Continuous Improvement assessment in Uganda and discuss its implications for HIS improvement in Uganda and elsewhere. Adele Waugaman will share experiences from leading a global expert group under the Health Data Collaborative supporting development and harmonization of global goods and also present about donor alignment efforts focused on harmonization of HIS and digital health tools and their application in East Africa.

**Expected Discussion**

Discussion is expected to focus on emerging trends and findings from the assessments and what the community can do collectively to address them.

**Timeliness of the Panel Topic**

This panel focuses on a topic of high priority for international organizations, donors, and national governments in their efforts to adopt and use digital health technologies to strengthen data-based decision-making necessary for improving both patient and population health outcomes. The Sustainable Development Goals (SDG) for health aim to improve access to and delivery of healthcare services for individual clients while addressing population health issues. In this context, international organizations, donors, and national governments in LMICs are prioritizing digital health interventions such as electronic health records (EHR). For instance, the digital health effort for Africa launched recently by the World Health Organization-Africa Region Office and the International Telecommunication Union emphasizes generic adoption of digital technologies in the national health system. Also, the Digital Regional East African Community Health Initiative is another large-scale effort to use health information technology to improve healthcare outcomes but is focused on the larger health system. Furthermore, the World Health Organization of the United Nations is in the process of publishing the first state of the art HIS status report by September 2019 that will highlight accomplishments, identify gaps and recommend appropriate interventions and tools to overcome gaps hindering progress towards the health SDGs.

**Organizer Statement**

All participants have agreed to take part on the panel.

**References**

Current Applications of Blockchain Technology in Biomedical Research and Healthcare

Didactic Panel

Tsung-Ting Kuo, PhD1, Amar Das, MD, PhD2, Kim Augustine, MS, BSBA, BFA3, Peng ("Dana") Zhang, PhD4, and Lucila Ohno-Machado, MD, PhD1,5

1 University of California San Diego, La Jolla, CA; 2 IBM Research, Cambridge, MA; 3 Deloitte, Rosslyn, VA; 4 Vanderbilt University Medical Center, Nashville, TN; 5 VA San Diego Healthcare System, La Jolla, CA

Abstract

In this 90-minute didactic panel, we will discuss the latest biomedical research and healthcare blockchain applications and implementations. Blockchain is a novel and emerging technology. However, its use in biomedical informatics is still emerging. The target audience includes medical informatics researchers, clinicians, health IT professionals, and leaders of healthcare organizations. The moderator and panelists, who are experts from academia and industry, will provide insights and representative use cases. We will discuss the utilization of blockchain to advance healthcare informatics, the use of blockchain in clinical research, the combination of blockchain and HL7’s Fast Healthcare Interoperability Resources (FHIR) to share clinical data, as well as the integration of blockchain with privacy-preserving predictive modeling. Our panel provides a unique and timely opportunity to see experts discussing myths, advantages, and potential concerns of adopting blockchain technology in healthcare and biomedical research.

Learning Objectives: Participants will understand the various use cases of blockchain in healthcare, clinical research, biomedical data sharing, and privacy-preserving modeling.

Keywords: Data Sharing, Interoperability and Health Information Exchange, Privacy and Security, Blockchain, Distributed Ledger Technology

Programmatic Theme: Clinical Informatics

Track: Application

Statement of Participation: All panelists have agreed to take part on the panel.

Introduction

Overview ( Moderator: Lucila Ohno-Machado, 10 minutes opening + closing discussion)
I will introduce the panelist and summarize the objectives of the discussion, placing it in the context of state-of-the-art medical blockchain applications and implementations, with a short introduction to the concepts being used in the panel. I will briefly list issues related to data integrity and the setup, design and pros/cons of using blockchain, with brief general examples. Panelists will discuss specific use cases, including blockchain use as an immutable ledger, smart contracts, the combination of blockchain and HL7’s Fast Healthcare Interoperability Resources (FHIR) to share clinical data, and the integration of blockchain with predictive modeling to preserve privacy.

Blockchain and Healthcare Informatics (Panelist: Kim Augustine, 20 minutes talk)
Blockchain technology has the potential to transform healthcare by increasing the security/privacy/ interoperability of health data. This presentation will discuss blockchain and its potential impact to healthcare informatics. The presentation will cover: 1) Results of a survey or trends regarding blockchain adoption and legislation. 2) Guidelines and considerations for organizations that are interested in adopting a blockchain solution. 3) Demonstration of Deloitte’s blockchain proof-of-concept, developed for managing healthcare credentials, and lessons learned.
Blockchain for Clinical Research (Panelist: Amar Das, 20 minutes talk)

Controlling data access and ensuring data integrity are central to the management of data collected in clinical research. Blockchain can meet these technical requirements with its immutable distributed ledger, formal verification protocols, and fine-grain data sharing. In this presentation, I will illustrate the use of the open-source Hyperledger blockchain technology in supporting two clinical research applications: dynamic consenting and clinical trials. In particular, I will discuss network setups for each use case, the design of smart contracts and the use of private channels for data privacy. I will contrast the use of blockchain technology to database solutions for clinical research, and discuss the advantages and disadvantages of each approach.

Blockchain Clinical Data Sharing on FHIR (Panelist: Peng ("Dana") Zhang, 20 minutes talk)

For collaborative clinical decision making, a secure and scalable way to share data is essential. However, conventional clinical data efforts are often siloed, which can create barriers to timely information exchange and delay effective treatment decision-making for patients. This presentation will introduce a conceptual framework called "FHIRChain" that leverages blockchain technology to facilitate clinical data sharing in the context of the technical requirements defined in the Office of the National Coordinator for Health Information Technology (ONC)’s “Shared Nationwide Interoperability Roadmap”. First, we provide analyses of these requirements and their implications for blockchain-based systems. Second, we present the FHIRChain architecture designed to meet those ONC requirements using FHIR standards for shared clinical data. Finally, we summarize key lessons learned from designing the architecture to demonstrate the value and limitations of blockchain technology’s applications in healthcare applications.

Blockchain-based privacy-preserving modeling (Panelist: Tsung-Ting Kuo, 20 minutes talk)

The protection of individual privacy is imperative for predictive modeling on healthcare data across institutions. However, current state-of-the-art privacy-preserving predictive modeling methods are mostly centralized, which may incur risks such as single-point-of-breach. Therefore, we leveraged blockchain, a distributed ledger technology that has been proposed for many healthcare applications, to decentralize the predictive model learning process, and therefore mitigate the risks of single-point-of-breach. Specifically, we adopted blockchain to store the partially trained models as transactions, and to update the models in a peer-to-peer way. Blockchain also enabled additional benefits, such as immutability and easy availability of the transactions. We designed an algorithm to integrate blockchain’s transaction metadata and privacy-preserving machine learning. Our algorithm also benefited from blockchain technology to address security and robustness weaknesses. The integration of blockchain and privacy technologies for cross-institution healthcare modeling can help minimize privacy risks and can support biomedical researchers in building and testing high-quality predictive models in a distributed, yet privacy-preserving, way.

Significance

Blockchain is an emerging technology. Considering the rising number of studies related to blockchain-based biomedical applications, our panel, composed of experts from academia and industry, will provide insights and describe uses cases in a way that is accessible by non-experts. Our panel provides a unique and timely opportunity to discuss myths, advantages, and potential concerns of adopting blockchain technology in healthcare to a wide and diverse audience that includes biomedical informatics researchers, clinicians, health IT professionals, and leaders of healthcare organizations.

Discussion Questions

1. What are the main commonalities and differences between data sharing solutions (e.g., the proposed FHIRChain for health information exchange, typical approaches for clinical data research networks)?
2. What concerns have organizations voiced regarding blockchain?
3. How would you prioritize blockchain features for adoption in your healthcare system?
Acknowledgements

LO-M and T-TK are funded by the National Institutes of Health (R01GM118609, R01HL136835, R13HG009072, K99/R00HG009680). T-TK is also supported by the UCSD Academic Senate Research Grant RG084150 and Travel Award TG087935. The content of this panel is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Veterans Health Administration.

References

HCD and HIT Modernization: Engagement of End Users in Health IT Modernization through Human-Centered Design

Maia Z. Laing, MBA1, Theresa Cullen, MD, MS2,3, Jennifer Shivers, BS, MFA2, Joseph Amlung, MPH2, Sabrina Fonseca, BA4
Office of the Chief Technology Officer, Immediate Office of the Secretary, U.S. Department of Health and Human Services, Washington, DC1, Global Health Informatics, Regenstrief Institute, Indianapolis, IN2, Indiana University School of Medicine, Indianapolis, IN3, PRTNRS (&Partners) Product Design and Development Studio4

Abstract
The HHS IHS HIT Modernization Project is sponsored by the U.S. Department of Health and Human Services Office of the Chief Technology Officer (HHS OCTO) in conjunction with the Indian Health Service (IHS) to provide an analysis and recommendations to HHS concerning the modernization of the IHS Health Information Technology (HIT) system (RPMS). This project is anticipated to be the first step of a multi-faceted project. Human-Centered Design (HCD) is an integral methodology for the evaluation of the current state and is being used to evaluate the current state of HIT across the I/T/U (Indian Health Service/tribally operated/urban programs) and to identify alternatives.

This panel will demonstrate the value of a human-centered research approach early in the software development lifecycle and will provide participants with methods, tools and recommendations for applying HCD in other health informatics projects. They will share the experience of applying HCD methodology to help answer “What does modernization look like from the end-user perspective?” The methods described were applied during IHS hospital and clinic site visits and during analysis sessions. Lessons learned, including appropriate tools to use for gathering, synthesis and analysis, will be shared.

Introduction and Background
Over the last several years, it has become common to apply human-centered design methods to the user interface design aspects of medical informatics projects1. Historically, these projects focus on reducing medical errors and on designing interfaces and user interactions with the system. More recently, key stakeholders have begun to request that human-centered approaches be applied to broader aspects of informatics projects, such as a Veterans Affairs project that evaluated the current use of a clinical information system in intensive care units3. This 2015 study broadened the use of HCD to include interviews, and found that the system existed in a “complex sociotechnical system” to encompass factors beyond the computer interface. In addition, similar HCD methods are being used to design global health technology solutions3. A human-centered approach is outlined in the US Government’s Digital Services Playbook4. This playbook outlines the importance of understanding the technology that people need and addressing the complete user experience. The IHS first used HCD methodology in 2004 to help design graphical user interfaces for a limited number of software applications in RPMS, and has been committed to this methodology when funding has been available. The HHS IHS HIT Modernization Project supported the use of integrated HCD methods from the inception of the project to assess what is working well, and more importantly, to understand system users’ challenges. The team employed an HCD approach designed to uncover technological issues, and the HCD methods highlighted challenges with the technology process and training. The panel will share the methods and processes used and the challenges of using a human-centered approach. They will also address the challenges of analyzing qualitative data with a team that is scattered geographically and will identify the process and methods the team used to analyze qualitative data. Results specific to RPMS, including the impact of a commitment to health equity impacts, HCD results, and lessons learned will also be shared.

Panel Description
Maia Laing will lead a discussion addressing how the team applied HCD to the information-gathering phase of the HHS IHS HIT Modernization Project. She will emphasize the importance of HCD methods in understanding the system user’s perspective and needs. The panel discussion will include an overview of the role of HCD within the U.S. Department of Health and Human Services, CMA, as well as in this current project. The decision to employ HCD methods and how this decision advances the study will be shared. The panel will describe the HCD methods used and will share the iterative process employed to refine methods and streamline the information gathering and analysis of qualitative data. Finally, key learnings will be shared with a discussion regarding how these methods can be applied to other health informatics projects.
Presenters

Overview of IHS Project -- How HCD Fits into the Project

Dr. Cullen will discuss the project goals and purposes, and how HCD fits in with the other project methods that are being used. She will share the early experience of using HCD in software development since 2004 within HIT in Indian Health Service, and lessons learned working cross-culturally. Dr. Cullen will also discuss the reasons behind embracing an HCD approach and the perceived and actual value of using an HCD approach in the discovery or fact-finding phase of the project. Finally, Dr. Cullen will address the balance between a design team and a technical assessment and how the team approached that challenge. Dr. Terry Cullen is a family physician who retired from the US Public Health Service as Rear Admiral in 2012 after leading multiple software development and deployment initiatives within the IHS. Between 2012 and 2015, Dr. Cullen worked as the chief medical information officer for the Veterans Health Administration.

HCD Project Approach and Methods

Jennifer Shivers will discuss the importance of aligning the methods with the context, team skills, and the phase of the project. She will share the HCD methods such as listening openingly\(^2\) and shadowing\(^5\) that were used to engage with system users in determining “What does modernization look like from the end-user perspective?” She will discuss and provide examples of the affinity mapping method\(^3,6,7\) used to analyze the findings from site visits. She will also discuss synthesis and how that process was used on the project. Finally, Jennifer will share information on how the team, who had little HCD experience, prepared and planned for executing the methods in a human-centered way and discuss how other teams might apply these types of methods to early phases of research. Ms. Shivers has a BS in computer science and a terminal master’s degree in human-centered design from Indiana University. In addition, she has previous experience teaching Human Computer Interaction and systems analysis classes at a university level.

HCD Tooling, Process, Findings and Next Steps

Joseph Amlung is a Business Analyst with the Global Health Informatics group at the Regenstrief Institute. He utilizes his background in public health to bridge the gap between research and practice; this involves designing and carrying out the process for data collection, along with orienting team members to the overall goals and the methods used to achieve them. Mr. Amlung works with the IHS HIT Modernization Project to enhance data collection by applying methods of human-centered design to team visits to American Indian/Alaska Natives (AI/AN) healthcare facilities. He will describe the processes used to capture, digest, and communicate findings and how those processes were iterated to make it easier to operate with a team that is not geographically co-located. Additionally, he will describe the information that was collected at these healthcare facilities and how it contributed to the overall goals of the HIT Modernization Project. Finally, Mr. Amlung will share the process for turning research into insights and actionable steps, and he will discuss findings to-date and the next steps for the research team.

Challenges and Lessons Learned

Sabrina Fonseca will discuss challenges in using HCD, training the team using these methods, and best practices for using human-centered methods. She will also discuss the lessons the team learned about group size, fitting the method to the context, making sure participants feel safe to share information. She will also provide learning points and recommendations on training for teams using HCD methods. Ms. Fonseca is the Director of Design Research at PRTNRS (&Partners). She has over 12 years of experience leading every stage of the user-centered design process, focusing on discovery, user research, strategy, conceptual design and iterative user testing. Working at leading design firms, Sabrina has uncovered people's needs, articulated problems, and broken down complicated tasks into easy, delightful experiences for clients such as Walgreens, Tuple Health, Planned Parenthood, Bayer, E*Trade, J.P. Morgan Chase, Nike, and others in the healthcare, enterprise, financial services, and non-profit sectors.
Discussion Questions

1. What are the benefits and challenges of using HCD methods in health informatics projects?
2. What was the process of learning how to incorporate HCD methods into a health informatics project?
3. What are the benefits, barriers and challenges of using HCD methods and research practices for more than application interface design?
4. How can barriers to using human-centered research approaches be reduced or removed?
5. What tooling or strategies can be used to increase the success of HCD?
6. What are some approaches for removing barriers for teams that are not co-located, yet want to analyze and synthesize together?
7. What are some of the challenges in researching and designing for complex systems including multiple user types performing complex team activities?

Justification

The increasing complexity of health care delivery as well as poorly designed health information technology solutions contributes to provider and health care team frustration, as well as inefficiency. Designing with the user in mind is not a novel concept, but it has historically been an afterthought in HIT. In this session, we will share HCD derived lessons learned (including tools and technologies) during an analysis of a currently deployed system in an underserved population that are easily integrated and extensible to others as organizations strive to ensure that HCD is part of their clinical informatics portfolio.

Panel Organizer Statement: All participants have agreed to take part in the panel and discuss the topics as outlined above.

References

**Decision Support for Genomically-Informed Cancer Therapy**

Funda Meric-Bernstam, MD¹, Matthew Rioth, MD, MS², Assaf Halevy, B.SC.³, Pete Stetson, MD, MA⁴

¹MD Anderson Cancer Center, Houston, Texas, ²University of Colorado, Denver, Colorado, ³2bPrecise, Pittsburgh, Pennsylvania, ⁴Memorial Sloan Kettering Cancer Center, New York, New York

**Abstract**

Next-generation sequencing is increasingly being used for clinical decision-making in oncology. There is a growing number of genomic testing platforms, genomically-selected trials, and rapidly evolving literature in genomic medicine and molecular oncology. Thus there is a tremendous need for decision support to help health care providers and clinician-investigators. Emerging needs include assistance in accessing, displaying, and interpreting genomic reports, appropriately determining actionability of genomic alterations and identifying approved and investigational therapeutic options. The learning objectives of this panel include determinants of actionability for genomic alterations, needs for decision support in annotations of functional implications of specific alterations and therapeutic implications, and resources and approaches for decision support for genomically-informed therapy.

**Panel Description**

Genomic testing with next-generation sequencing is widely utilized for patient care both in the community and academic centers. The growing size of genomic test panels has increased the opportunity for identifying “actionable” genomic alterations targetable with approved or investigational agents. However, it also increases the burden on health care providers to access and accurately interpret genomic test results within the time constraints of a busy clinical workflow. This challenge is compounded by the rapidly evolving literature and changing availability of treatment options.

Key principles of decision support include determination of actionability, with a focus of therapeutic actionability for treatment with approved and investigational agents.¹, ² Access to and visualization of genomic data and incorporation of reports into the electronic health records (EHR) remain specific problems. The large number of emerging testing platforms add challenges in interpretation of reports taking into account differences in the platforms. Tumor-normal testing present different problems as it pertains to interpretation of potential alterations of germline origin.

Interpretation of actionability heavily relies on interpretation of functional impact of genomic alterations.¹, ² This is being addressed by a variety of local and programmatic efforts that lack harmonization. The interpretation of variants of unknown significance is an area of specific concern, both due to concerns that their actionability may be misperceived and because they represent a communication challenge as the knowledgebase evolves.

Although there have been several recent approval for therapies linked to a genomic biomarker, still most genomically-relevant therapies are only available through clinical trials. Clinical trial matching increasingly relies on genomic markers. Thus annotation of gene-drug associations and maintaining trial availability up to date as well as developing methods to alert treating physicians and providers of trial options are important challenges to deliver on the promise of genomically-informed therapy.

**Funda Meric-Bernstam, MD (moderator and organizer)** will review the clinical needs in decision support as well as selected challenges in genomically-informed therapy. She will review a framework of actionability of genomic alterations and potential approaches to developing decision support tools. She will present a use case of local implementation of enterprise-wide Precision Oncology Decision Support.³, ⁴ She will present approaches for building and maintaining a precision oncology knowledgebase, rationale for tools such as on-demand precision oncology decision support reports, clinical trial alerts and clinical trial search engines. Metrics for success in decision support will be discussed.

**Pete Stetson, MD MA (panelist)** will review solutions for ordering and display of genomic testing reports integration in the EHR with actionable annotations.³ He will review the usage patterns of advanced visualization tools by oncologists which augment traditional molecular pathology reports. He will report on methods for incorporating genomic results within provider documentation workflows. He will discuss implementation and governance issues around sharing genomic testing reports with patients in a patient portal. He will report on progress on translational informatics tools to support genomic-driven clinical trials enrollment.
Matthew Rioth MD MS (panelist) will discuss methods of integration of external molecular profiling results within an electronic health record and relevant data standards. A specific case report of this will be presented. He will discuss the handling of “variants of unknown significance (VUS)” as well as their research and potential clinical relevance.

Assaf Halevy, B.SC (panelist) will review the vendor perspective on integration of multiple testing platforms, content development, terminology harmonization, point-of-care workflow integration, and integration of genomic data with relevant clinical data for actionable display. He will present design considerations for both somatic tumor testing, as well as pharmacogenomic testing. He will discuss challenges and opportunities in integrating with medical center clinical systems which otherwise do not have existing genomic decisions support solutions.

Questions to be considered include:
1) What are the best practices for integrating into the EHR workflow for the ordering providers?
2) How do we define actionability?
3) How do clinicians and scientists determine the functional impact of genomic alterations?
4) What are the informatics challenges to creating and maintaining a gene-drug association knowledgebase?
5) What are the best practices for tracking changes in actionability such as VUS that later become important can be acted upon?
6) What are best practices for annotating genomically-relevant trials?
7) How do we provide decision support for genomically-matched clinical trial enrollment?
8) What are the best options for engaging patients and displaying genomic test results and therapeutic interpretations?

Topic Rationale: Genomic testing is increasingly being utilized in oncology. Oncology is an especially important use case for decision support as genomic testing is now a mainstream tool and there is a large number of biomarker linked drug approvals and therapeutic options emerging every day. Thus this panel on Precision Oncology is very timely. The topic presents challenges in multiple informatics domains including bioinformatics, information retrieval, clinical research informatics, clinical decision support and consumer health informatics, and thus it is likely to be of interest to the informatics community at large.

Statement of Agreement to Participate: All panelists approved this submission and have agreed to participate.

Panel Participant Details:
Funda Meric-Bernstam, MD. Chair, Department Investigational Cancer Therapeutics, Medical Director, Institute of Personalized Cancer Therapeutics, University of Texas MD Anderson Cancer Center

Pete Stetson, MD, MA. Chief Health Informatics Officer, Deputy Physician-in-Chief, Office of the Physician-in-Chief, Memorial Sloan Kettering Cancer Center

Matthew Rioth, MD. Assistant Professor, Medicine and Bioinformatics, Director of Clinical Cancer Informatics, University of Colorado Cancer Center

Assaf Halevy, B.SC. Founder and CEO, 2bPrecise

Conflict of interests will be provided.

References
Unlocking the potential of real-world evidence (RWE) in oncology – technology- and informatics-enabled solutions for clinical care, precision medicine, and discovery

Robert S. Miller, MD\(^1\), Wendy S. Rubinstein, MD, PhD\(^1\), Andrew D. Norden, MD, MPH, MBA\(^2\), Shawn M. Sweeney, PhD\(^3\), William S. Dalton, MD, PhD\(^4\)

\(^1\)CancerLinQ\(^\text{®} \) LLC, American Society of Clinical Oncology, Alexandria, VA, \(^2\)COTA, Inc., New York, NY, \(^3\)AACR Project GENIE, American Association for Cancer Research, Philadelphia, PA, \(^4\)Oncology Research Information Exchange Network, Tampa, FL

Abstract

This didactic panel will describe how real-world data, traditionally defined as those data derived from a source other than a traditional clinical trial, can generate real-world evidence and can be used to facilitate research and inform clinical care. Focusing on the field of oncology and the collection, aggregation, and transformation of cancer data predominantly but not exclusively contained within electronic health records, panelists representing four platforms currently delivering solutions – CancerLinQ\(^\text{®} \), COTA, GENIE, and ORIEN – will provide their perspectives on what it will take to improve cancer data interoperability and make meaningful progress in knowledge- and data-sharing to benefit patients. The panelists will discuss the need for alignment on data standards, challenges around curating and incorporating rich clinical data sourced from unstructured documents and free-text, and the particular data needs of precision oncology.

Panel description and issues to be examined

A model predicated on the principles of a rapid-learning system (RLS) has been advocated as an antidote for the unsustainable cost, unevenness of quality, and lack of patient-centricity plaguing U.S. healthcare, and the field of oncology has been suggested as a compelling first choice to develop and test such a system.\(^1\) A technology-enabled RLS could be foundational in managing oncology’s contemporary challenges – the explosion in the amount of genomic and other molecular data needed to understand cancer development and progression in individual patients and populations, the precipitously rising costs of cancer therapies, persistently low participation in clinical trials by adult cancer patients (3-5% at most), and continued provider- and system-level administrative and reporting burdens that impact data collection for downstream analytics.

While a number of non-profit, academic, and commercial entities have made great strides in building some of the framework and key components of an oncology RLS, a universal, truly interoperable system is not yet a reality. Without question, there remain immense challenges in developing a system that can seamlessly and efficiently aggregate data, analyze it, and provide guidance back to clinicians and insights to researchers, while meeting patients’ needs rapidly enough to be effective. Nonetheless, there is a growing recognition that today’s care delivery systems and the traditional research and drug development framework built on the paradigm of Phase I-IV clinical trials are simply inadequate to close current knowledge gaps in genomics (and other “-omics”), outcomes, and cost/value. Optimizing the capture and use of “real-world data (RWD)” and “real-world evidence (RWE),” variably defined but generally viewed as that arising from sources other than traditional clinical trials,\(^2\) will be required. In this context, RWD/RWE will predominantly focus on electronic health record (EHR) data, since these are the primary transactional systems in use, but is meant also to include insurance claims data, prescription refill data, data from cancer registries, death data, genomics databases, data from patient-reported outcomes collection and wearable devices, and other sources.

This panel will examine the central theme of how oncology RWD is collected, aggregated, and understood and how oncology RWE is then generated. The panelists will review existing technology- and informatics-enabled solutions. All of the panelists have deep and varied domain expertise in data issues and have been directly involved in the creation and operation of some of the solutions. Using examples from their own work, the panelists will help address some broader themes:
Description of panelists’ topics and affiliations

**Wendy S. Rubinstein, MD, PhD (CancerLinQ)**
CancerLinQ is a big-data, health technology platform created and operated by the American Society of Clinical Oncology (ASCO) for quality improvement and discovery. This talk will address some of the approaches used by ASCO and CancerLinQ to improve cancer data interoperability, including work to identify a set of minimal common oncology data elements. Mechanisms of data aggregation and normalization, including integrating and extracting value from unstructured data, will be described.

**Andrew D. Norden, MD, MPH, MBA (COTA)**
COTA is a cancer-focused data and analytics company that applies technology and human experts to create structured oncology datasets from electronic health record (EHR) data for diverse purposes including clinical quality/value improvement and research. This talk will address best practices and quality assurance requirements for combining technology and manual curation of data. A novel, clinically driven patient cohort generation mechanism known as the COTA Nodal Address (CNA), that creates homogenous groups of patients, will be described.

**Shawn M. Sweeney, PhD (AACR Project GENIE)**
The American Association for Cancer Research (AACR) Project GENIE is an international data-sharing consortium focused on generating an evidence base for precision cancer medicine by integrating clinical cancer genomic data with clinical outcomes data from 19 institutions internationally. This talk will describe the need for clinically-annotated, genomics datasets, and describe the numerous challenges in gathering such data including level of annotation/metadata and use of standards.

**William S. Dalton, MD, PhD (ORIEN)**
The Oncology Research Information Exchange Network (ORIEN) is a research alliance between 19 US-based comprehensive cancer centers utilizing a common protocol for tracking patient molecular, clinical and epidemiological data and following them throughout their lifetime. This talk will describe how prospective cohort surveillance of patients may be a solution to accelerate the development of new therapies and improve patient enrollment on clinical trials.

**Robert S. Miller, MD (moderator)**
The moderator will frame the 4 talks in the context of basic informatics principles around data standards and secondary use of data. The critical need for new solutions and models to modernize cancer research, development, and care delivery, especially to make it more patient-centric, will be emphasized.

Importance of panel topic and intended audience

Assessing the value and limitations of RWE has become a necessary competency for clinicians, researchers, regulators, payers, and policymakers alike, particularly in a data-intensive field like oncology. However, the specifics of data capture, aggregation, and transformation, challenges with curating unstructured content, and role for common data elements, remain murky for most unfamiliar with the field. This panel will focus on the needs of oncology clinicians and researchers, who are the primary audience. Oncologists faced with the challenges of managing the care of cancer patients using today’s EHRs, given widespread dissatisfaction and usability issues, nonetheless recognize
the potential of data sources beyond clinical trials. Using practical examples from their own platforms, the panelists will demonstrate how RWE can reveal patterns of care and can inform the comparative assessment of treatment toxicities and costs. Researchers see the limitations of the traditional clinical trial paradigm but remain uncertain of the validity of RWE and how it could complement prospectively-collected trial data. Focusing primarily on the integration of molecular data with phenotypic data, the panelists will demonstrate how RWE databases can be used for hypothesis generation and how innovations such as pragmatic trial designs using an RWE control arm or prospective cohort surveillance can be valuable.

Discussion questions for audience participation

1. How can clinicians be incentivized to improve data capture at the point of care without adding to their current administrative and reporting burdens?
2. How comfortable will treating clinicians be making decisions for the care of individual patients based on RWE (i.e., observational data from large cohorts of like patients) and not guidelines traditionally based on results from clinical trial?
3. Under what circumstances can manual curation of datasets be eliminated or drastically curtailed, while maintaining adequate analytic value?
4. What is the optimal level of granularity for a minimum oncology dataset – only the highest level common data elements applicable to all cancer patients vs. a data specification that is more expansive (i.e., designed for individual anatomic and/or molecular subtypes)?
5. How can oncology clinicians and researchers begin to formulate an approach for engaging their own healthcare organizations and their EHR vendors and sequencing labs (as customers) in supporting better collection of structured data at the point of data generation, particularly molecular data, for better downstream analytics.
6. How do we create a national (global) infrastructure to share data from all networks involved in collecting and studying patient data that will enable all stakeholders to access and learn from the data to better meet patient needs?

ALL CO-AUTHORS HAVE AGREED TO ATTEND THE 2019 AMIA ANNUAL SYMPOSIUM AND TO PARTICIPATE ON THIS PANEL.

References:

Curating EHR data in the All of Us Research Program
Karthik Natarajan, PhD; Robert Carroll, PhD; Thomas R. Campion, Jr., PhD; Joan M. Grand, RN, MS; Shyam Visweswaran, MD, PhD

1Department of Biomedical Informatics, Columbia University, New York, NY
2Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
3Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, NY
4MITRE Corporation, 202 Burlington Road, Bedford, MA
5Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Abstract
The All of Us (AoU) Research Program has embarked on an ambitious effort to integrate data from a diverse number of sources to create a resource that will enable biomedical research in the next decade and beyond. One large data source is EHR data. We will present the experiences of collecting, submitting, and curating data from various AoU enrollment sites to the Data and Research CEnter (DRC). The topics covered in this panel will include: the goals of the curation pipeline within the DRC; the experience with participant matching at both the DRC and enrollment sites; the data mapping efforts to the Observational Medical Outcomes Partnership (OMOP) vocabulary; and the development of metadata to support data quality.

Introduction
The national All of Us (AoU) Research Program is a historic effort to gather a rich set of data from 1 million people living in the United States to create an unprecedented resource that will enable biomedical research in the next decade and beyond1,2. The goal of the program is to revolutionize how disease is prevented and treated based on individual differences in lifestyle, environment and genetics. One key source of data that is being integrated at an unprecedented scale are the electronic health records (EHRs) of participants. Currently, enrollment sites include over 30 Healthcare Provider Organizations (HPOs) and Federally Qualified Health Centers (FQHCs) that are enrolling participants across the nation and transferring EHR data to the Data and Research Center (DRC). The DRC acquires, organizes and provides secure access to what will be one of the world’s largest biomedical data resource that includes one of the largest consolidation of disparate EHRs into a single central repository for precision medicine research.

Evaluating EHR data quality for secondary use has become an active area of research in the past decade3-6. Predominantly, these evaluations have been done on single data sources and the coordination required to evaluate the data quality of individual sites within a network has additional challenges. In order to create a data resource with consistent data quality for research, all sites within a network need to adhere to a set of standard mapping conventions. Similar to other networks, the DRC is developing guidelines to represent clinical data, such as lab results and medications. Two primary differences between other networks and AoU is the scale at which EHR data is being aggregated and the development of a suite of computational research tools for large-scale analyses of the data in a cloud platform.
The DRC is developing a set of data quality checks to ensure the creation of a data resource that researchers can trust and supports a broad spectrum of research uses. In order to support this objective, the DRC must work in partnership with the EHR data partners (HPOs and FQHCs) to establish curated data submissions to the DRC. This panel will focus on the data curation process of EHR data from both the DRC and HPO/FQHC perspective.

**Topic**
The panel represents informaticians from various HPO/FQHCs and the DRC. They will share their experience in the collection and submission of EHR data, as well as the data quality issues that arise. Each panelist will present processes implemented, successes, and challenges faced with extracting and submitting data. We will discuss the data curation pipeline within the DRC and how it supports EHR data quality checks both locally at sites as well as within the DRC. The panel will conduct a structured conversation on the following topics:

- Goals of the curation pipeline within the DRC
- Experience with participant matching algorithms employed at the DRC and HPOs
- Processes around identifying and improving data mapping to the OMOP vocabulary
- Development of metadata to support data quality

**Panel Participants**

**Robert Carroll (Moderator) - Vanderbilt University Medical Center**
Dr. Carroll is a Research Assistant Professor of Biomedical Informatics whose research focus is on research using EHR data in conjunction with DNA biobanks to study human health and disease. In the field of phenotyping with EHR data, he has developed and applied algorithms and machine learning methods to predict disease state and drug response. He has participated in the eMERGE Network and other consortia, implementing and designing methods intended for EHR-based cohorts and more traditional observational cohorts, in particular studying the ability of such methods to be applied across disparate data sets. His recent efforts have focussed on managing and harmonizing data sets, including directing the transition of the Vanderbilt EHR research repositories, the Research Derivative and the Synthetic Derivative, to the OMOP common data model in part to support the joining of multiple EHRs of data. He has been deeply involved with the DRC and is a co-lead of the data curation effort which is responsible for the creation of a research data set spanning many domains including survey, EHR, and genomics data for the *AoU*.

**Karthik Natarajan – Columbia University Medical Center**
Dr. Natarajan is an Assistant Professor of Biomedical Informatics at Columbia University Medical Center. His focus is on developing scalable information retrieval and data mining methods for clinical data. He is a co-director of the Biomedical Informatics Resource in the CTSA and manages the clinical data warehouse at Columbia University’s Medical Center. Part of that work involves leading efforts to implement the OMOP data model using clinical and administrative data from multiple sources. He is the co-lead of the EHR operations committee for the *AoU* and works closely with the DRC, and is also involved in data modeling efforts of the PCORI-funded New York City-Clinical Data Research Network (NYC-CDRN). Dr. Natarajan will describe the curation process of EHR data within the DRC.
Thomas Campion – Weill Cornell Medical Center
Dr. Campion is an Assistant Professor of Healthcare Policy and Research at Weill Cornell Medical Center. He is the director of research informatics at Weill Cornell where his group helps investigators obtain data from EHR systems, collect novel measures that do not exist in EHR systems, and integrate data from disparate sources. Dr. Campion’s team is responsible for sharing Weill Cornell EHR data with a number of multi-institutional initiatives, including AoU, the PCORI-funded NYC-CDRN, NCATS-funded Accrual to Clinical Trials (ACT) network, and TriNetX. For AoU and other large studies, he has led efforts to support study coordinators with participant identity and relationship management to enable subsequent extraction of EHR data.

Shyam Visweswaran – University of Pittsburgh
Dr. Visweswaran is an Associate Professor of Biomedical Informatics. His research interests include the application of artificial intelligence and machine learning with a specific focus on learning EHR systems and computerized clinical decision support, precision medicine and personalized modeling. Dr. Visweswaran serves as the Director of the Biomedical Informatics Core for the University of Pittsburgh Clinical and Translational Science Institute (CTSI), as a PD/PI for the All of Us Pennsylvania HPO, and as the Data Harmonization lead for the ACT network. He also directs the development and implementation of a EHR research data warehouse at the University of Pittsburgh. Dr. Visweswaran will share the challenges and successes the All of Us Pennsylvania HPO has had with submitting EHR data to the DRC.

Joan M. Grand – MITRE Corporation
Ms. Grand is the project lead for the AoU FQHC Phase II Project. As a registered nurse, Joannie brings more than 25 years of extensive clinical and administrative experience in the ambulatory, critical care and emergency medicine domain. While at MITRE, she has been at the forefront of efforts to leverage data, evidence-based practice and analytic methods to improve the health and health care of the nation while building trusting relationships with various stakeholders to create catalyzing partnerships. Ms. Grand will provide an overview of the FQHCs’ engagement and share their experiences within AoU.

References
Real World Experiences with Patient-Facing APIs for Interoperability, Access, and Use of Electronic Health Data

Aaron B. Neinstein, MD;¹ Julia Adler-Milstein, PhD;¹ William Morris, MD;² Anil Sethi, MS;³ Elise Anthony, JD⁴

¹University of California, San Francisco, San Francisco, California; ²Cleveland Clinic, Cleveland, Ohio; ³Citizen, San Francisco, California; ⁴Office of the National Coordinator for Health Information Technology, Washington, D.C.

Abstract

Application programming interfaces (APIs) underpin the progression from data to knowledge to action in so many walks of life. In healthcare, APIs have recently received national attention as Medicare and Medicaid began requiring doctors and hospitals to provide their patients with digital access to their health data through patient-facing APIs. Now the Office of the National Coordinator for Health IT has released groundbreaking rules for standardized APIs as a cornerstone of national interoperability and patient access. As healthcare’s API era launches, our panel will present four key perspectives on patient-facing APIs—academic research, frontline health systems, app innovators and developers, and national policy—and strategies and priority use cases in the real world today. We will present findings of a just-completed national study of leading health systems and their uses of patient-facing APIs, including best practices, early challenges, and potential policy implications, followed by initial research on patients’ usage of those APIs, a deeper look at one leading health system’s strategy and experience, a leading app developer’s perspective, and ONC’s major policy proposal this year using APIs to advance patient access and interoperability. This panel provides an eagerly anticipated first look at real-world usage and impacts of patient-facing APIs.

Panel Description

Numerous industries outside healthcare have proven that robust application programming interfaces (APIs) are necessary infrastructure to move from data to knowledge to action. Electronic health records began as and often continue to be health data repositories and billing systems for doctors and hospitals. However, their transformative value rests on the ability to share health data from these silos to the many, varied stakeholders in the learning health system which would benefit from more seamless data access and use. Key among these stakeholders are patients.

In 2015, the Office of the National Coordinator for Health Information Technology (ONC) and Centers for Medicare & Medicaid Services (CMS) proposed to launch a potential democratization of electronic health information exchange by requiring the use of patient-facing APIs. CMS and ONC proposed that patients gain access to their electronic health data stored in EHRs through open APIs that would allow patients to view, download, integrate, and share their health data wherever they chose: a family caregiver, an emergency department, a community clinic, a third-party digital health app, etc. CMS and ONC originally required that these patient-facing APIs take effect on January 1, 2017, but delayed the requirement to January 1, 2019.

Thus, after much anticipation, patient-facing APIs are now live in many places in the United States. However, little information has been published about their real-world usage and impacts. Making the technology available does not necessarily ensure the concomitant culture, business drivers, and workflows needed to use it well. This panel, comprised of national leaders in their respective perspectives, will provide a first look at early real-world experiences with patient-facing APIs from four diverse perspectives—academic research, health systems, national policy, and consumer-facing applications—to provide an important initial synthesis of the promise and real world experience to date. Following the opening presentations, the moderator will lead a conversation among panelists and audience with questions intended to stimulate additional learnings and perspectives on patient-facing APIs.

Aaron Neinstein, MD (moderator and panelist) – National Landscape and Early Experiences

Dr. Neinstein will introduce the panel and topic. He will open the session by presenting data from a recently-completed national study, funded by The Commonwealth Fund, of how health systems are implementing patient-facing APIs. The study involved interviews with ten leading health systems to discover lessons learned from how they are implementing and using patient-facing APIs, including drivers, strategies, use cases, associated organizational
infrastructure, governance, policies, available applications, and processes. The findings reveal variability among these leading health systems with respect to their emphasis on and use of patient-facing APIs.

Julia Adler-Milstein, PhD (panelist) – Patient-Facing API Usage Data

Dr. Adler-Milstein will present new empirical findings on how often patients use their smartphones to connect and download their EHR data from their health systems through Fast Healthcare Interoperability Resources (FHIR®)-enabled APIs. The research asked 12 health systems to share anonymously the numbers and trends of unique and cumulative individuals who linked their smartphones and downloaded their data over a nine-month period (3/29/2018-12/31/2018).

William Morris, MD (panelist) – Health System Perspective

Dr. Morris will describe the experience of a leading national health system and its perspective on strategy, use cases, implementation, challenges, and best practices in using patient-facing APIs.

Anil Sethi, MS (panelist) – App Innovator and Developer Perspective

Mr. Sethi will next discuss the experience of an app innovator and developer, its perspective on the state of patient-facing APIs, and how patient-facing APIs might meet its strategy and use cases to provide access and interoperability to patients.

Elise Anthony, JD (panelist) – Federal Policy Perspective

Ms. Anthony will describe the ONC’s draft regulations on APIs and information blocking, as a major nationwide development to advance patient-facing APIs, patient access, and interoperability.

Questions for Panel and Audience Discussion

- As we look at some of the expectations around patient-facing APIs – driving consumer or patient access and use of health data, driving emergence of a third-party application ecosystem, enabling better health information exchange – how have patient-facing APIs lived up to these and other expectations? Where have results unfolded as expected?
- What have been the unexpected lessons learned in this first year of the national rollout?
- Where are we discovering gaps between what we hoped would happen and what appears to be happening?
- In what ways are patient-facing APIs the same as or different from other types of APIs used in healthcare? What are the implications for how each stakeholder approaches its API strategy?
- What are the predominant current use cases for patient-facing APIs? Do you think they will remain relatively unchanged over the next 2-3 years, or will they evolve or expand?
- How important is inclusion of write APIs alongside read APIs, and why? When might that happen on a significant scale?
- How have patient-facing APIs, and the patient access, use and sharing of health information they enable, changed clinical care?
- What are the remaining policy gaps with respect to patient-facing APIs?

Rationale for Topic and Panel

Multiple forces pushed patient-facing APIs to center stage in 2019 as a national strategy and tool for access and interoperability with patients. CMS and ONC began requiring Medicare and Medicaid clinicians and hospitals to use EHRs with open APIs certified to the 2015 Edition to enable patients nationally to view, download, and transmit their health data. Apple Health used patient-facing APIs to roll out availability to over 190 million patients across America. ONC and CMS proposed new regulations to use standardized APIs to promote national interoperability and access as well as standardized policies to prevent information blocking with APIs. This panel will present early research and insights about health system implementations of patient-facing APIs, including a timely and unique first look at strategies, use cases, policies, challenges and lessons already learned, and issues to come. This will highlight the variability among these leading health systems with respect to their strategic emphasis on patient-facing APIs.
Panel Participants:

- **Aaron Neinstein, MD.** Director, Clinical Informatics, Center for Digital Health Innovation, and Assistant Professor, Department of Medicine, University of California, San Francisco.
- **Julia Adler-Milstein, PhD.** Director, Center for Clinical Informatics & Improvement Research, and Associate Professor, Department of Medicine, University of California, San Francisco.
- **William Morris, MD.** Associate Chief Information Officer, Cleveland Clinic.
- **Anil Sethi, MS.** Founder and Chief Executive Officer, Citizen.
- **Elise Anthony, JD.** Executive Director, Office of Policy, Office of the National Coordinator for Health Information Technology.

Statement of Agreement to Participate:

All panelists approved this proposal and agreed to participate.

Learning Objectives:

- Learn and discuss key perspectives on patient-facing APIs—academic research, frontline health systems, app innovators and developers, and national policy—and strategies and priority use cases in the real world today.
- Learn the key findings of a just-completed national study of leading health systems and their uses of patient-facing APIs, including best practices, early challenges and pitfalls, and potential policy implications.

Presentation Type:

- Panel (traditional)

Programmatic Theme:

- Consumer Health Informatics

Key Words:

- Interoperability and Health Information Exchange
- Patient Engagement
- Patient / Person Generated Health Data (patient reported outcomes).

Additional key words:

- Application Programming Interfaces (APIs)
- Patient Access
- Mobile Access

Tracks:

- Policy
- Applications
Informatics-Enabled Learning Health Systems: Strategies for Success from Four Academic Medical Centers

Eric G Poon, MD MPH FACMI1,2, Charles P Friedman, PhD FACMI3, Philip RO Payne, PhD FACMI4,5, Michael J Pencina, PhD2,6, Kevin B Johnson, MD MS FACMI7,8

1Duke University Health System, Durham NC; 2Duke University School of Medicine, Durham, NC; 3University of Michigan Medical School, Ann Arbor, MI; 4Washington University School of Medicine, St. Louis, MO; 5Washington University McKelvey School of Engineering, St. Louis, MO; 6Duke Clinical Research Institute, Durham, NC; 7Vanderbilt University Medical Center, Nashville, TN; 8Vanderbilt University School of Medicine, Nashville, TN

Abstract

The confluence of EHR adoption, advances in data science, and intensified pressures to achieve the quadruple aim has provided fertile ground for the development and spread of Learning Health Systems (LHS), which the Institute of Medicine has defined as systems in which ‘science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience.’ Informaticians will need to play a key role in these ecosystems, but many socio-technical barriers may stand in the pathway to success. This panel will showcase the experience of four academic medical centers that have been early developers of LHS. Panelists will describe strategies employed and investments made to overcome these barriers and sustain the success of LHS.

By the end of this panel session, participants should be able to i) describe structures, methods, and activities used commonly in learning health systems; ii) articulate the success factors for functional and sustainable learning-health systems, and iii) formulate strategies to optimize the set-up of learning-health system initiatives at their local institutions.

Significance

The wide-spread adoption of EHRs in US hospitals and ambulatory practices coincides with the intensified focus on the quadruple aim to optimize patient outcomes, improve patient experience, contain cost of care, and prevent clinician burnout1. Academic medical centers are staring at the eye of this perfect storm as public2,3 and private payors4 demand provider organizations to compete based on value for patients. Academic medical centers, as bastions for research, innovation, and teaching, need to develop strategies that leverage their core expertise to continuously improve operational excellence and define new paradigms of value-based care5. The construct of Learning Health System (LHS) has been proposed and implemented in various forms across several premier academic medical centers to address these challenges.

An LHS is defined by the Institute of Medicine (now named National Academy of Medicine) as a system in which ‘science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience.’6 As conceptualized by Dr. Friedman, one of the panelists, they are made up of multiple ‘virtuous cycles’ of study and improvement, each with an afferent (blue) and an efferent (red) arm:

![Figure 1. The Learning Health System Cycle](image)

122
Various informatics-related disciplines and their underlying methodological approaches can empower both the afferent and efferent arms of any LHS. For example:

i) Data amassed in the EHR and other clinical information systems enables data-driven inquiry at the local, regional and national levels

ii) Analytic tools allow clinicians, leaders, and researchers to ask question at the population level at the point of care

iii) Recent data science advances allow diverse end-users to synthesize data to make accurate and actionable predictions on which patients might benefit the most from timely and customized interventions and/or wellness promotion strategies

iv) Consumer informatics allows clinicians to see the full picture of their patients and extend the reach of interventions into the daily lives of patients and their loved ones, while also engaging patient as active members of the care team

v) Biostatistical, epidemiological, data science, and qualitative methods allow organizations to measure the impact of programs and continuously optimize or refine them.

In short, LHS embodies the theme of the 2019 AMIA Annual Symposium ‘From Data to Knowledge to Action’. However, adoption of learning health systems is still at its infancy, and informaticians must partner with many disciplines and stakeholders for LHS to be successful. Approaches for LHS vary greatly, and there may be many pathways to success. This panel will allow its participants to learn from early adopters of LHS and apply their lessons learned to their local communities.

Description of Panel

The 90-minute didactic panel session will consist of 4 brief 10-minute presentations by the luminaries in the field of informatics and learning health systems. Each presentation will be followed by 4 minutes for questions and answers. The final 30 minutes will be devoted a moderated discussion between the panel members and the attendees. By the end of this session, participants should be able to achieve the following learning objectives:

i) Describe structures, methods, and activities employed commonly in learning health systems

ii) Articulate the success factors for functional and sustainable learning-health systems

iii) Formulate strategies to optimize the set-up of learning-health system initiatives at their local institutions

The didactic panel will be moderated by Dr. Eric Poon, Chief Health Information Officer at Duke Health. After stating the learning objectives of the panel session, Dr. Poon will propose a working definition of learning health systems, and outline how the confluence of US EHR adoption, advances in data science, and intensified pressures to achieve the quadruple aim provides fertile ground for the spread of learning health systems activities. He will highlight some common socio-technical challenges that might hinder the development of thriving learning health ecosystems.

Dr. Charles Friedman, Chair of the Department of Learning Health Sciences at the University of Michigan School of Medicine, will propose a framework for LHS success and provide an overview of the general state of LHS adoption in the US and European countries. He will discuss the goals, structure, and research agenda of Department of Learning Health Science at University of Michigan. He will describe the bottom-up and middle-out strategy employed to form collaborations with other medical departments on LHS projects, including their campus-wide “LHS Collaboratory” and other collaborations at the state level. Dr. Friedman will also describe ongoing efforts to imbue LHS approaches through the Clinical Decision Support Committee and Clinical Intelligence Committee at Michigan Medicine.

Dr. Philip Payne, Founding Director for the Institute for Informatics at the Washington University School of Medicine in St. Louis, will describe efforts to advance Health IT and Biomedical Informatics platforms to deliver data as a renewable resource, translate data assets into actionable knowledge, and return ensuing insights to both clinicians and patients. He will review approaches for achieving timely and ready access to data assets through the creation and management of appropriate data repositories and sharing platforms that address the privacy and security concerns in a scalable manner. He will then describe efforts to promulgate advanced data analytics and knowledge management techniques across the research and care delivery enterprise. He will particularly emphasize the design and evaluation of data-driven interventions that involve the use of clinical decision support, guideline delivery
systems, population health management platforms, and tailored patient-level or population-level messaging at-scale in order to address the quadruple-aim.

**Dr. Michael Pencina, Vice-Dean for Data Science and Information Technology at the Duke University School of Medicine**, will discuss efforts at Duke Health to create learning health units that empower clinicians, operational leaders, clinical researchers, data scientists, methodologists, IT professionals, and change management practitioners to collaborate on LHS activities on a consistent platform of shared resources. This approach, facilitated by a "Connector" organization, allows efficient access to curated data and methodological expertise, while allowing LHS activities to scale across all clinical departments and service lines at Duke Health under shared governance. He will also highlight how the suite of consultation services, data pipelines, self-service data tools, on-demand burst computing resources in a secure platform, a common Information Asset Repository, and workflow integration approaches support the full-cycle of LHS activities at Duke.

**Dr. Kevin Johnson, Chair of the Department of Biomedical Informatics at Vanderbilt University Medical Center**, will discuss how Vanderbilt is developing a pipeline to connect their successful EHR-based CDS, ordersets, and forms modalities with their LHS to conduct pragmatic effectiveness trials. That pipeline takes advantage of their CTSA Studio Program to shape research ideas into hybrid designs. These studios inform the next step, EHR-powered research workshops, designed to assess the feasibility of implementing the study design within the constraints of the EHR or the existing clinician-EHR workflow. Dr. Johnson will showcase a number of examples, including smoking cessation, pediatric obesity management, and genome-informed pharmacotherapy.

**Discussion Questions:**

1) What are the common gaps in skillsets panelists and attendees have encountered in their organizations while setting up a functional learning health system? What have been some approaches to bridge those gaps?

2) What creative ways have been used to leverage existing resources and identify new funding sources to support learning health activities at your organization and/or your communities?

3) How might one go about proving the return-on-investment from learning health system activities?

4) If the panelists and attendees were given the opportunity to set up another learning health system, what would they do differently?

5) What opportunities exist for separate learning health systems to band together as a system of systems to accelerate knowledge generation and magnify impact?

**Statement of Agreement:** All panelists named on the authorship list have agreed to take part in this panel.

**References**


7. https://medicine.umich.edu/sites/default/files/2014_12_08-Friedman-ION%20LHS.pdf (last accessed 3-12-2019)


Implementing e-PROs into Clinical Practice

Robert S. Rudin, PhD, SM1, Cynthia LeRouge, PhD2, Danielle C. Lavallee, PharmD, PhD3, Madhu Reddy, PhD, FACMI4, Anuj K. Dalal, MD5

1 RAND Corporation, Boston, MA; 2Florida International University; 3University of Washington; 4Northwestern University; 5Brigham and Women’s Hospital and Harvard Medical School

Abstract

Patient-reported outcomes (PRO) are widely used in research studies, but thus far have limited use in routine clinical care. Despite the potential to make care more patient-centered, improve the accuracy of symptom recognition, facilitate shared decision making, and improve patient outcomes, health systems struggle to implement PROs in a sustainable way that generates value for patients and clinicians. Electronic health records (EHR), patient portals, and mobile Health apps have the potential to collect electronic PROs (e-PROs) more seamlessly, but best practices regarding implementation have not been established. In this panel, four presenters will describe emerging lessons from experiences implementing e-PROs in the clinical setting. Topics include methods for PRO visualization; designing and implementing a practice models that collects e-PROs via smartphone app for patients; integrating e-PROs collected from apps into the EHR for clinicians; and governance considerations. Attendees will learn the current state of knowledge for effective e-PRO implementation as well as key challenges and future directions.

General Description

Patient-reported outcomes (PRO) allow patients to systematically report their symptoms, functional status, quality-of-life and other relevant aspects of their health. Symptoms reported directly by patients have been shown to more accurately reflect patient health status compared with reports by clinicians. Systematic collection of PROs can facilitate the detection of health problems, provide numerical comparisons with other patients, inform treatment decisions, and assess the quality of healthcare delivery. Collecting these data systematically also has implications for population-based health care delivery.

Although PROs are widely used in the context of research studies and some subspecialty settings, they have not been widely adopted as part of routine clinical care. Electronic health records (EHR) have the potential to enable more routine use of e-PROs, such as when collected online, via patient portals, and in-person via kiosks and tablet computers. However, little is known about how to optimally implement e-PROs for their broader and more routine use for clinical care.

This panel presents e-PRO implementation from multiple perspectives at two different institutions: UW Medicine (Seattle, WA) and Brigham and Women’s Hospital (Boston, MA). The panel will discuss a range of topics, such as methods for e-PRO visualizations within the EHR, the design and implementation of a practice model for collecting condition-specific e-PROs via a smartphone app for patients with asthma, infrastructure requirements for integrating ePROs into routine care, and governance considerations for maintaining an e-PRO program. Collectively these presentations will provide a broad overview of the current state of e-PRO use in routine care and will offer suggestions of how to advance the science of e-PRO implementation.

Moderator: Madhu Reddy, PhD, FACMI is a Professor at Northwestern University. He researches topics at the intersection of medical informatics and computer-supported cooperative work (CSCW), with a focus on improving care delivery through use of collaborative health information technologies. He has recently published an evaluation of ePROs used in routine care. He will introduce the topic and each panel member, and lead the discussion.

Panel Members

Cynthia LeRouge, PhD is an Associate Professor in Florida International University and holds an adjunct position at the University of Washington School of Public Health. Topic: Visualizing PROs for providers. As collection of PROs increases across healthcare settings there is a need to develop e-PRO tools that incorporate effective mechanisms (e.g.,
visualizations) for reporting PROs to providers. Many clinicians perceive a PRO system as a potentially valuable addition to augment their patient care process if preferences for use and visualizations are met and if PRO workflow is streamlined. Literature and our action research findings focused on designing ePRO visualizations for providers reveal many commonalities among clinician perspectives of e-PRO data visualization as well as nuances. While a moderate degree of visualization customization may be needed for a system-wide deployment of e-PROs to meet clinician expectations of PRO usefulness and ease of use, some general considerations for e-PRO reports exist. This panelist will review an integrated e-PRO visualization framework that references three critical design domains related to content (i.e. what type of data is provided), visualization design (i.e., how data is visually represented), and functionality (i.e., which capabilities are provided) of e-PRO tools. Review of this framework will showcase common targets for ePRO visualization design as well as key points of customization needed to support various specialties and provider preferences.

Robert S. Rudin, PhD is an Information Scientist at the RAND Corporation. Topic: Designing clinically integrated PROs as part of a practice model: case study of asthma. PROs are beginning to be integrated into routine care in the context of a patient encounter, but they also have the potential to facilitate between-visit monitoring for chronic conditions such as asthma. Current guidelines recommend that clinicians monitor asthma symptoms, but few do so routinely. This panelist will present an asthma symptom monitoring practice model and mobile health (mHealth) app that collects PROs weekly between visits. The practice model and app were designed using user-centered design methods. Results of a 6-month feasibility study in subspecialty care shows high patient retention and adherence and minimal burden on the clinic. The panelist will discuss preliminary designs for adapting this intervention to the primary care setting; considerations for long term scaling and sustainability; and barriers and facilitators to implementation.

Anuj K. Dalal, MD is an Assistant Professor at Harvard Medical School, and Associate Physician at Brigham and Women’s Hospital. Topic: Leveraging the EHR infrastructure to implement a PRO-based practice model for a health system. Effective implementation of informatics interventions require integration with EHRs and other electronic infrastructure. Third party applications offer more flexibility regarding their design and use, but are more challenging to integrate into existing workflow and the EHR. This panelist will discuss how Brigham and Women’s Hospital supports various options for integrating third-party applications into its EHR (Epic Systems, Inc.), and describes how the above example (presented by Rudin) will be implemented. He will describe technical capabilities (e.g., FHIR, web services), workflows considerations (e.g., triage protocols), integration strategies, and governance considerations.

Danielle Lavallee, PharmD, PhD is a Research Associate Professor at the University of Washington. Topic: PRO Governance. The increasing interest in using PROs for clinical care is one example where governance can support clinical implementation efforts while keeping an eye towards needs and priorities of the system. PRO Governance ensures formalized oversight to PRO implementation including decisions related to selection, implementation, analysis, and use of PROs within and across the healthcare system. This panelist will present one example of approaching systemwide governance to manage competing demands for IT resources across different areas of clinical care. She will address the role governance plays in balancing different aspects of PRO implementation. This includes creating community and culture, supporting policy and procedure development, managing stewardship of IT resources, and keeping in mind needs for quality and contractual reporting.

Why topic is timely

The promise of using PROs to make care more patient-centric has so far been elusive. In general, direct patient input regarding their conditions is still not collected systematically or used as part of routine care. Now that EHRs have been widely adopted, collecting PROs electronically is more feasible: in theory, patients should be able to enter their data via online portals or mHealth apps directly into the EHR, simplifying workflows and allowing data to be collected outside of visits. Similarly, clinicians should be able to access PRO data more efficiently and leverage visualizations for comparative and longitudinal assessment of PRO data. As health systems increasingly are pressured to become more responsive to patients’ needs, there is an urgent need to understand how to implement PROs more sustainably in ways that will realize the potential benefits at the individual patient and population levels.

Anticipated audience
Anyone interested in learning how to make care more patient-centric using technology would be interested in attending this panel. Individuals with a focus on PROs, clinically integrated mHealth, health IT governance, and/or patient-centered practice models would be particularly likely to attend this panel.

**Discussion questions**

We anticipate a lively discussion with the audience. The moderator will encourage the audience to share their experience with implementing e-PROs. Possible questions include:

- What are the key benefits of incorporating e-PROs into routine care? How do these benefits differ by context and condition? What conditions are most promising for e-PRO collection?
- Which types of medical specialties would likely be the consumers of e-PROs for clinical care?
- How can e-PROs be presented to providers in ways that makes them clinically useful? What is the range of visualization options?
- How do health systems balance e-PRO standardization while also customizing to specific clinical practice needs?
- What role can mHealth have in e-PRO collection before, during, after, and in-between visits?
- What are workflow barriers and facilitators for clinicians to using e-PROs during visits? How does this differ by clinical profession?
- What are the technical barriers and facilitators to implementing e-PROs in a health system’s EHR?
- What are governance best practices and challenges to implementing e-PROs?
- What incentives are driving adoption (or non-adoption) of e-PROs in routine care?

**Statement from panel organizer** All participants have agreed to take part on the panel.

**Acknowledgements**

RSR was supported by the Agency for Healthcare Research and Quality grant #1R21HS023960. RSR and AKD were supported by the Agency for Healthcare Research and Quality grant #1R18HS026432. CL and DCL were supported by the Agency for Healthcare Research and Quality grant #R01HS023785

**References**

Sync for Science™ – A Standards-Based Approach to Data Sharing by Patients for Clinical Care and Research

Anita Samarth¹, Kevin Chaney, MGS², Joshua C. Mandel, MD³–⁴, Spencer L. SooHoo, PhD⁵
¹Clinovations GovHealth, Washington, DC; ²Office of the National Coordinator for Health Information Technology, Washington, DC; ³Microsoft Healthcare; ⁴Harvard Medical School Department of Biomedical Informatics, Boston MA; Cedars-Sinai Medical Center, Los Angeles, CA⁵

Abstract

In 2015, the Office of the National Coordinator for Health Information Technology (ONC) and the National Institutes of Health (NIH), in partnership with Harvard Medical School, announced the launch of Sync for Science (S4S) a national collaboration to accelerate adoption of application programming interfaces (APIs) that provide individuals a simplified, scalable, and secure way to share their electronic health record (EHR) data. S4S provides software tools, implementation support, documentation, and a convening platform to foster consistent, standards-based patient API access across diverse EHR systems. Initially tested to support EHR data sharing with the All of Us Research Program under the Precision Medicine Initiative®, S4S is a significant step towards developing a health system that provides patients with greater control over their EHR data, including aggregation of data across multiple provider sites, and sharing it securely and efficiently with other services, providers and research programs via mobile applications (apps). This effort leveraged participation from ONC, NIH, participating developers, researchers, providers, and participants insight regarding barriers to implementing and using S4S, which can inform other similar endeavors. While S4S is testing sharing of EHR data for precision medicine research, the standards powering it can be leveraged for other uses. This panel will discuss key lessons learned from the S4S pilot thus far, and opportunities for leveraging S4S and standards-based APIs to support health care and research.

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act, enacted as part of the American Recovery and Reinvestment Act of 2009, helped spur broad-based adoption of certified electronic health record (EHR) technology across the United States. While this has enabled digitization of health information and relevant health care services, it has not been without its challenges. Seamless sharing of electronic health information with patients, caregivers, or amongst providers is not yet ubiquitous. To address issues with access to health information, Congress enacted the 21st Century Cures Act (Cures Act) in 2016. The Cures Act strengthens ONC’s mandate to improve interoperability and, in particular, directs ONC to “engage in rulemaking to drive patient, clinician, and payer access to clinical data by advancing proposals related to application programming interfaces (APIs)”.

Standards-based application programming interfaces (APIs) are expected to enable individual access to their EHR data for any purpose of their choice. Excitingly, one specific purpose, such as participating in a research study, is currently being piloted by the All of Us Research Program which is part of NIH’s Precision Medicine Initiative® (PMI). The emerging app ecosystem supported by the consumer technology industry- could be an important catalyst for changing how individuals manage, share, and use their health data for their own health and for research²⁻³.

Sync for Science™ and the All of Us Research Program

To advance the goals of the PMI, ONC, in partnership with NIH has launched several activities to assist making this ambitious vision a reality⁴. One of those, Sync for Science (S4S)⁵, seeks to demonstrate how participant data donation can be scaled across large-scale research efforts such as All of Us, using standards-based API technology. Conducted in collaboration with Harvard Medical School and four leading EHR developers, S4S is being piloted at a variety of clinical sites across the country – allowing individuals to share their electronic health data specifically for All of Us. Further, S4S uses technical specifications that allow individuals to share their EHR data with a research study of their choice or care entity, also aligns with the requirements of the Centers for Medicare & Medicaid Services (CMS) under
the Promoting Interoperability Program (MU3), and, ONC’s recent notice of proposed rulemaking, which proposes requiring health IT developers to support API-enabled services. These policy drivers will continue fueling the expansion of S4S and other standards-based APIs. S4S provides important insights into the technical integration, roll-out, and use of S4S - informing our understanding of implementing an API allowing real-world data transfer for health care research. This panel seeks to provide insight into how API technology for participant data donation can be scaled across large-scale research efforts, based on the early experiences of S4S.

Panel Objectives and Presenters

The aim of this panel is to provide an overview of S4S, which is piloting under the Precision Medicine Initiative’s cornerstone program, the All of Us Research Program. Key lessons learned, areas for improvement, emerging challenges facing the field, and priorities for future work will be discussed by the diverse stakeholders involved. The panel brings together the informatics, research, and policy experts leading this work and will highlight pilot experiences from Cedars-Sinai Health System, Duke University Medical Center, Partners HealthCare, and University of Missouri Health Care and S4S lessons-learned from APIs within Allscripts, Cerner, eClinicalWorks, and Epic EHRs.

Ms. Anita Samarth, the CEO and Co-Founder of Clinovations GovHealth, will introduce and moderate the session.

Mr. Kevin Chaney, a Senior Program Manager at ONC, will provide an overview of ONC’s priorities and their relation to S4S to enable data sharing, lessons learned and emerging themes from S4S to improve standards-based API use.

Dr. Joshua C. Mandel, Chief Architect at Microsoft Healthcare, and project lead for S4S, will discuss the specifications, technical integration, and roll-out of S4S in conjunction with the All of Us Research Program.

Dr. Spencer SooHoo will discuss how the pilot is proceeding at Cedars-Sinai Health System, and describe how it is using standards-based APIs within the Epic EHR that are leveraged to enable health data access for S4S and the Apple Health mobile app. Dr. SooHoo will discuss the testing processes with S4S, All of Us, Epic, and Apple Health.

Panel Discussion Questions

- Is the health IT ecosystem keeping pace with the utilization of APIs? Are efforts by consumer technology companies spurring a more robust app ecosystem?
- What technical and policy needs or gaps have been identified through your work limiting the adoption of S4S and other, standards-based APIs? Are there limitations on the kinds of data sent?
- What are the barriers and facilitators for the pilot sites participating in S4S? Are providers familiar with APIs?
- Are the pilot sites interested in continuing their association with All of Us, and/or expanding to other research activities that might be relevant?
- How might ONC’s proposed rules around “APIs without special effort” influence shifts in research?

Panel Learning Objectives

- Participants will have a better understanding of NIH’s All of Us Research Program and how Sync for Science can support participant access and sharing of electronic health record (EHR) data for research.
- Participants will learn about the technical specifications and workflow of implementing S4S, using a standards-based application programming interface (API).
- Participants will learn what challenges and barriers may impact the use of a standards-based API, like S4S, for promoting data access and sharing for research.
- Participants will learn about emerging data types that could be leveraged by S4S for access and sharing by researchers.
- Participants will learn how S4S aligns with federal policy initiatives and emerging challenges that could impede the use of similar APIs.
Conclusion

The *All of Us* Research Program represents an inflection point for conducting research at a scale never seen before, driving a new generation of evidence. Sync for Science plays a critical role in allowing individuals to easily share their health data with any research or care entity of their choice. Therefore, early insights into the development, use, and improvement of open standards for data sharing is critical to the success of S4S – especially as policy drivers are expected to further implementation and usage of APIs.

Statement of Participation

Each of the panelists and the moderator have confirmed that they will participate if this submission is accepted, at the assigned timeslot during the Annual Symposium.

References

FHiring up Evidence in CDS: Mobilizing Knowledge for Computable Guidelines in Patient Care

Lisa M. Schilling, MD, MPSH1, Robert A. Greenes, MD, PhD2, Brian S. Alper, MD, MSPH, FAAFP3, Bryn Rhodes4, Maria Michaels, MBA, PMP5

1Professor of Medicine, University of Colorado, Aurora, Colorado, 2Ira A. Fulton Chair and Professor in Biomedical Informatics, Arizona State University, and Professor of Biomedical Informatics, Mayo Clinic, Scottsdale, Arizona; 3VP of Innovations and EBM Development, EBSCO Health, Ipswich, MA, USA, 4Co-Chair, HL7 Clinical Decision Support Work Group; CTO, Dynamic Content Group, 5Centers for Disease Control and Prevention, Atlanta, GA, USA

Session Type: Didactic Panel

Abstract

This panel brings together informatics experts in key domains necessary to leverage current and future technologies to efficiently translate evidence into action by making evidence and clinical practice guidelines computable and represented in a standard, interoperable format. There is a recognized lag of about 17 years in the translation of evidence to practice. The Internet, new data sources (such as ‘omics profiles and home sensor data), and new methods (such as causal inference and predictive analytics), and sources of evidence (such as ‘real-world evidence’) will lead to ever faster and broader evidence generation. Computable clinical guidelines, even when they exist, have been difficult to integrate into care processes due to a lack of pragmatism and patient-centeredness, among other factors. The panel will address challenges of guideline creation and adoption in the digital age, and the extent to which the availability of data and ways of acquiring and delivering knowledge provide new opportunities for integration and use of best-practice knowledge. The panel will discuss innovation in FHIR to transfer information regarding evidence (EBMonFHIR), clinical practice guidelines (CPGonFHIR), various types of computable knowledge and standards for biomedical knowledge (the Mobilizing Computable Biomedical Knowledge (MCBK) initiative), and redesigned processes for clinical practice guidelines.

Introduction and Background

The transformation of evidence to action is key for optimal health care delivery and outcomes. Evidence-based practice may best be implemented through the use of clinical decision support tools (CDS) and other EHR-assisted pathways. As methods of evidence generation continue to expand due to new analytic methods (e.g., causal inference, predictive analytics) and growing data availability (e.g., ‘omics, sensors, imaging), while simultaneously becoming more tailored (e.g., personalized medicine) there are even greater opportunities to support clinicians and patients in the practice of evidence-informed care. The reputed 17-year gap between medical knowledge and evidence-based medical practice will continue to exist if computerized methods are not used to assist decision-making.(1)(2)

‘Knowledge application – the act of getting evidence into practice – is fraught with challenges, some of which are specific to given environments (3), and others systematic, such as the efficient translation of guidelines into machine-readable formats. (4)(5). Embedding clinical practice guidelines into CDS system has resulted in improved clinical adherence to recommendations.(6)(7) Despite a decade of innovation, there remains limited progress in the routine creation of computable clinical guidelines and evidence (8), which are still an afterthought of their narrative format. This panel will address recent progress in methods to solve these challenges with use of improved guideline creation methods, FHIR, and other data standards for representing clinical practice guidelines and evidence knowledge to ensure computability, re-usability, and the ability to share.
Outline of Topics
Welcome and Introductions: Lisa Schilling, MD, MSPH will provide a brief overview of the panel, and introduce the presenters.

Panelist Presentations (in order of presentation)

Maria Michaels, MBA, PMP is Public Health Advisor for the Deputy Director for Public Health Science and Surveillance at the Centers for Disease Control and Prevention (CDC). Ms. Michaels leads a multi-stakeholder initiative, Adapting Clinical Guidelines for the Digital Age, which aims to redesign the guideline development and implementation process to include all perspective from the outset of guideline development as well as use standards-based approaches to connect the evidence with the tools needed to apply guidelines in patient care. Part of this initiative includes an HL7 implementation guide on representing clinical practice guideline recommendations in FHIR (CPGonFHIR). Ms. Michaels will discuss the challenges of developing and implementing clinical guidelines which are often likened to a ‘game of telephone’ with inconsistencies and inaccuracies in translation, and a lag time that lead to an average of 17 years before scientific evidence is routinely applied to patient care. Redesigning the process to develop standards-based computable guidelines and include downstream perspectives during guideline development can help minimize translational issues. Ms. Michaels will discuss methods being developed to ensure that guideline developers and implementers are aware of the entire cycle for applying evidence to patient care and the use of standards such as Fast Healthcare Interoperability Resources (FHIR) throughout the cycle.

Robert A. Greenes, MD, PhD, is Ira A Fulton Chair and Professor in Biomedical Informatics at Arizona State University. Dr. Greenes’ research has been in the areas of clinical decision support, in terms of models and approaches for decision making, the knowledge representation to support it, and its clinical application and validation. He has also been active in the promulgation of standards and fostering of group collaborative work, particularly in knowledge management. He is a member of the Steering Committee of the Mobilizing Computable Biomedical Knowledge (MCBK) initiative, and co-chairs its Standards Workgroup. He will discuss this initiative’s goal to enable the curation, dissemination, and application of medical knowledge at a global scale. The management of computable biomedical knowledge will be essential to the success of Precision Medicine, the achievement of high-functioning Learning Health Systems, and more generally to the advancement of biomedical science.

Brian S. Alper, MD, MSPH, FAAFP is board certified in clinical informatics and is the project lead for the EBMonFHIR project, an HL7-approved project. The project’s goal is to develop a standard for computable expression of evidence, a critical underpinning of interoperability across the Evidence-Based Medicine (EBM) community of practice and with other types of computable biomedical knowledge. He will demonstrate precisely how FHIR-based resources for evidence assertions and statistics makes this type of biomedical knowledge computable. Dr. Alper has extensive leadership and practical solutions development across evidence-based medicine including founding DynaMed (a point-of-care clinical reference), defining Systematic Literature Surveillance (a methodology to keep current with best evidence), advancing guideline development (Guidelines International Network working groups), and multiple clinical decision support developments (including Patient-Centered Clinical Decision Support Learning Network).

Bryn Rhodes is Chief Technology Officer of Dynamic Content Group. Mr. Rhodes is co-chair of the HL7 CDS Work Group, the lead author of the HL7 CQL standard, and editor of the HL7 FHIR Clinical Reasoning module, as well as several other quality improvement-related HL7 standards. Mr. Rhodes is also a project champion in the Informatics Stream of the Adapting Clinical Guidelines for the Digital Age initiative and is actively involved in work sponsored by the CDC and ONC to create computable recommendations for the CDC’s Opioid Prescribing

Figure: The Plan-Do-Study-Act cycle: applying evidence to patient care
Guideline. Mr. Rhodes will introduce the CPGonFHIR project. The project aims to collect and document experience from the implementation community with using the FHIR Clinical Reasoning module resources to build computable clinical practice guideline recommendations, helping to apply those recommendations in patient care more easily, quickly, accurately, and consistently.

Summation (10 minutes): Dr. Schilling will provide closing comments and take questions from the audience.

All panelists have agreed to participate in this panel, which will provide important updates in this emerging field.

Educational Objectives
After participating in this session, the learner should be able to:

- Describe current progress in Clinical Practice Guideline (CPG) design and formalization to support computability
- Describe methods for generation of clinical decision support knowledge
- Explain alternative approaches to delivering knowledge at the point of care
- Describe efforts underway and approaches to sharing of best-practice knowledge
- Explain how biomedical evidence (research findings) can be expressed in a computable form

Intended Audience
This session is open to clinicians, informaticians, researchers, and policy makers who have an interest in the lifecycle of evidence generation to clinical guidelines to computerized clinical decision support.

List of Discussion Questions
1. Rethinking what is ‘evidence’: how can we use other data sources and ‘real-world evidence’ to inform clinical guideline updates and/or clinical decision support, and should we?
2. What are some current barriers, especially at the clinical organization level, to fully implementing standards such as FHIR?
3. What factors prevent sharing of knowledge artifacts (such as re-usable decision support rules, quality measures, and other tools) and how can we most effectively address those issues?
4. What are the challenges to implementing standards such as EBMonFHIR and CPGonFHIR?

References
Abstract

Biomedical informatics is at a crossroads with respect to the translation of innovation. The center of innovation and implementation in health IT has shifted from its historic origin in academia to industry. Applied researchers must become familiar with a new context, goals, processes and vocabulary in order to be able to translate their research ideas into practical products and services. Most researchers and business entrepreneurs share a common goal: to make a meaningful difference in people’s lives. But, to do so in the commercial space, academicians must keep several principles in mind, such as: addressing a real problem in a meaningful way; focusing on the customer; considering business development strategies up front; leveraging the core skills and competencies of academia and businesses in a complementary way; planning for scaling the innovation; respecting and protecting intellectual property; considering multiple pathways to commercialization; being realistic about valuation; leveraging a variety of sources for capital; and understanding their institution’s technology management and transfer processes. This panel, composed of successful academic and business entrepreneurs, will discuss a framework that can help attendees in going from research idea to initial public offering through a structured, easy-to-understand process.

Introduction

Biomedical informatics is at a crossroads with respect to the translation of innovation. The center of innovation and implementation in health IT has shifted from its historic origin in academia to industry. Applied researchers must become familiar with a new context, goals, processes and vocabulary in order to be able to translate their research ideas into practical products and services. This panel will provide an overview of how research ideas can be commercialized, and offer an easy-to-understand framework that attendees can apply.

Historically, the academic community has produced many significant and groundbreaking advances in informatics and health IT. For the most part, it did so by partnering with or integrating in local healthcare institutions, identifying relevant problems, and solving them with innovative basic and applied research. Sometimes, these solutions were disseminated and adopted beyond their origin. However, it was not uncommon for academic informatics organizations to function like loosely federated islands of innovation with little translation to the broader context of health IT. They also only used local, idiosyncratic processes for translating research ideas into production.

Panel description

The advent and growth of the health IT industry, the adoption of health IT at scale, as well as the relatively recent entry of large commercial players into the health IT space have fundamentally changed innovation in health IT. The center of gravity for health IT innovation has now shifted from academia and theory to industry and practice. Academic informaticians are slowly starting to pivot to innovating within electronic health record (EHR) systems. Newer technologies such as Fast Healthcare Interoperability Resources (FHIR) are facilitating this trend. At the same time, many academicians continue to pursue standalone research projects based on local needs.

What makes the current situation challenging is that the context, goals, processes and vocabulary that informaticians must understand in order to effectively translate their research ideas into practical products and services was typically not part of their training and/or experience. Added to this challenge are the sometimes significant hurdles in administrative red tape and opaque processes at their institutions, making the commercialization of research difficult if not impossible. How can researchers become more successful at translating their ideas into commercial applications?

Most researchers and business entrepreneurs intrinsically share a common goal: to make a meaningful difference in people’s lives. While their respective incentives differ, this common ground can become the basis of successful collaborations. A few considerations are important for academicians as they consider pathways to commercialize their ideas:
● **Address a real problem in a meaningful way**: Both researchers and business entrepreneurs must create value through their work. Projects that will not have significant impact on workflow, healthcare delivery, or patient outcomes, or meaningfully improve on status quo in the industry, will excite neither peer review panels nor investors.

● **Focus on the customer**: Solutions produced both in academia as well as in business must ultimately satisfy a customer need and demand. If that demand does not exist or cannot be stimulated, there is no market for the resulting product.

● **Consider business development strategies up front**: Many researchers spend years developing and testing their ideas, and often consider only fairly late in the development cycle how to translate their work into a commercial product/service. This behavior pattern often results in disappointment due to a variety of factors, such as competitive pressures, shifts in the technology landscape or changed market demand. Much better is a research and development strategy in which translation is a key consideration from the very beginning.

● **Leverage the core skills and competencies of academia and businesses in a complementary way**: Many academicians are very good at generating innovative ideas, but not as capable in translating those ideas into robust production applications. Successful businesses, on the other hand, tend to develop products/services that can stand up to the challenges of daily use. Productization by capable companies is one of the most direct routes from research idea to impactful product.

● **Planning for scaling the innovation**: An innovation that cannot be scaled for widespread adoption will not succeed. New technologies like the Apple Health Record coupled with the ubiquity of the iPhone means that innovations deployed through this platform could rapidly be scaled to 100 million users in the US—something completely out of reach for academia.

● **Respect and protect intellectual property (IP)**: Many researchers are justifiably proud of the IP they develop, but IP never exists in a vacuum. Other entities may own or have rights to or interests in the same or similar IP. Sometimes, partners contribute their respective IP in order to create something that is greater than the sum of its parts.

● **Consider diverse pathways to commercialization**: Very often, academics primarily focus on creating a spinoff company as a pathway to commercialization. However, there are other ways of leveraging IP, such as licensing to a business partner.

● **Be realistic about valuation**: One of the challenges with IP produced through research is that it is often difficult to value. Many universities spend inordinate amounts of time negotiating the potential value of an innovation and how any returns should be split. Very often, those negotiations lead nowhere, especially when the university has unrealistic expectations of financial returns. In the end, 50% of $0 is still $0.

● **Leverage a variety of sources for capital**: Researchers and business entrepreneurs can draw on a variety of sometimes complementary sources of capital. However, many researchers focus solely on research grants to fund their efforts. There are other options, however, such as SBIR/STTR grants and foundation awards. On the business side, angel investors and venture capitalists can provide capital for investment.

● **Educate yourself about your institution’s technology management and transfer processes**: Many universities have fairly complicated technology commercialization processes. One of the first things faculty must do is understand the commercialization process from their institution’s perspective, starting with the invention disclosure and ending with licensing/spinoff procedures.

The panel, composed of academics and business entrepreneurs who have brought a broad range of innovations to the market, will provide both academic and business perspectives on how to translate applied research to commercial use. It is targeted at biomedical informatics researchers and developers who produce innovations that are intended or appropriate for application in real-world clinical, administrative, research or educational environments. The topic is timely and relevant for a number of reasons, including the need to produce more impactful innovations that make a significant difference in practice, and the declining feasibility and usefulness of the concept-grant-paper cycle.

**Panelist Biographies**

**Titus Schleyer, DMD, PhD**, is Professor of Biomedical Informatics in the Department of Medicine, Indiana University School of Medicine, as well as a Research Scientist at the Center for Biomedical Informatics at the Regenstrief Institute. As a former programmer, he has developed several health IT production implementations over the course of his career, and recently refocused his research entirely on applied work that is implemented and evaluated in practice. His FHIR-based Chest Pain App was the winner of the inaugural AMIA Pitch IT Competition in 2018, and he is currently working with Intelligent Medical Objects in commercializing a derivative innovation.
Frank Naeymi-Rad, PhD, is the Chairman of the Board and co-founder of Intelligent Medical Objects, Inc. where he leads a team of informaticists, clinicians, and innovators in delivering industry-leading informatics services using medical terminology for electronic medical records (EMR) systems. Prior to founding IMO, Dr. Naeymi-Rad was co-founder of Intelligent Medical Systems, Inc., which was sold to Glaxo in 1994. His thought leadership in clinical informatics led to nomination for the Oracle Corporation Smithsonian Award in the category of Medicine. He will present a step-by-step journey showing his move from teaching computer science, AI and innovations with medical students and faculties to transactions with private equity firms. Dr. Naeymi-Rad will cover steps in his journey from seed money from friends and family, Angels, VC, Investment Bankers, Private Equity and strategize about his “secret sauce.”

Barbara Rapchak is the Vice President of Academic Innovation for Intelligent Medical Objects (IMO) and the Academic Innovation Accelerator (AI²). She is the founder of Leap of Faith Technologies, Inc., a digital health company, and eMedonline, a patented software-as-service platform for mobile medication therapy management that improves adherence and facilitates care transition and disease management, resulting in reduced healthcare costs for providers and payers, and enhanced drug use for pharmaceutical manufacturers. She has led sponsored research and technology development for the National Institutes of Health (NIH), collaborating with hospitals, universities, major corporations, and government agencies on a wide range of healthcare projects, and has been very successful in the NIH Small Business Innovation Research (SBIR) program. She will be the moderator for the panel.

Patricia Weber, DrPH, is a Program Director with the Small Business Innovation Research (SBIR) Development Center at the National Cancer Institute (NCI), National Institutes of Health. She manages a diverse portfolio of grants and contracts awarded by the NCI with an emphasis on digital health and health IT. She works with small businesses at all stages of the process from pre-application, award, post-award, as well as NCI program initiatives that assist with follow-on funding to commercialize NCI-funded technologies. Dr. Weber will discuss the NCI/NIH grant application process with a focus on the research scope of both early and later stage projects, the review process and interacting with program staff to get the early help that applicants need to be successful. In addition, she will discuss post-award I-Corps and commercialization assistance programs along with NCI specific initiatives such as the Investor Initiatives program.

Craig Feied, MD, is Chief Strategy Officer for the Microsoft Health Solutions Group and Professor of Emergency Medicine in the Georgetown University School of Medicine. Dr. Feied was a cofounder of the Informatics Section of the American College of Emergency Medicine. He was the first emergency physician in the United States to hold a position as Director of Informatics, and first to head an institute of medical informatics. He is a Diplomate of the American Board of Emergency Medicine, as well as a fellow of the American College of Emergency Physicians, the American Academy of Emergency Medicine, and the American College of Phlebology. Dr. Feied developed a unique practical software architecture ("Azyxxi") for the rapid integration, organization, display, and mining of all data present in any medical setting—including data from all legacy systems, across many regions—in real time. In 2006 the Azyxxi system was acquired by Microsoft Corporation as the enterprise core of a new Health Solutions Group, and in 2008 the system was renamed "Amalga."

All panelists have agreed to participate in this panel.
Ten Years of CTSA Outputs and Moving Forward: How We Succeeded, What We Learned, and Moving Forward in the Open Science Environment

Juliane Schneider\textsuperscript{1}; Kristi Holmes\textsuperscript{2}; Marc Ciriello\textsuperscript{1}; Griffin Weber\textsuperscript{1}; Diane Keogh\textsuperscript{3}
\textsuperscript{1}Harvard Medical School, Boston MA; \textsuperscript{2}Northwestern University Feinberg School of Medicine, Chicago, IL; \textsuperscript{3}i2b2 tranSMART Foundation, Boston MA

Abstract

This panel will follow the stories of three informatics tools or methods developed by CTSAs to look back on the successes and failures of ten years supporting clinical and translational science. The stories will highlight successes where they were perhaps not expected, lessons learned about tool development and outreach, how sustainability is supported, or why a great product never got widely adopted. The discussion will then turn to the current environment of open science, and how the changes in the research lifecycle could have impacted the success of tools that were ahead of their time, or made formerly widely used tools obsolete, and how future development needs to take into account open science best practices to maximize data and software reusability and funding requirements for data management. A discussion will elicit ideas and opinions from other stakeholders in the CTSA program, and how we can use the lessons of the past to ensure success in the next five-year cycle.

Description

The CTSA program was founded in 2006, with over 60 sites across the United States\textsuperscript{1}. Over the years, these sites have developed programs and tools to support the accomplishments of clinical and translational researchers, with targeted tools in response to needs within their local institutions and through collaborative efforts aimed at benefiting bioscience on a national level. Some tools have been widely adopted, some have not.

Since 2006, the scientific research cycle has gone from siloed, invisible processes to one that is increasingly working towards openness. This idea of openness pervades every step of the cycle, from the tools used, to how the data is gathered, stored and described, to the publication of results and attributions. How do the tools that have been developed, especially in the early years of the CTSA program, fit into the quickly changing research environment, and how does this new environment affect current efforts? What have we learned from successes and from the good ideas that were not widely adopted?

With ten years behind them, and the next five year cycle beginning, it is the time for the CTSAs to look back on successes, failures, and how they are going to incorporate the open science environment that is profoundly changing the research lifecycle. A discussion including all stakeholders - CTSA members, researchers and those interested in open science should take place in order to move forward with a clear idea of what is needed to succeed. We expect the discussion of past efforts will inform the discussion of the future.

This panel will review the successes and near-successes of some of the CTSA products in the context of the change to an open science model. The discussion will center around three CTSA-developed tools/methods and their success or failure, and how those results may have been different in the new era of open science. The stories of the tools presented will provide the context for a lively discussion on effective support of clinical and translational science, the lessons learned on outreach, implementation, technology, and collaboration over ten years, and how it all fits into the new open science environment.
Panel

All participants have agreed to participate in this session. -- Juliane Schneider, panel organizer

Kristi Holmes, PhD. Northwestern University Feinberg School of Medicine. Chicago, IL. Topic: Researcher Networking as a Model for Open Science Discovery, Communication, and Sustainability. Lessons Learned from the VIVO Collaboration.

Diane Keogh. i2b2 tranSMART Foundation. Boston MA Topic: i2b2 tranSMART Foundation - Creating a 501©3 non-profit Foundation as a model to support sustainability and community engagement.


Marc Ciriello. Harvard Medical School. Boston, MA. Topic: Moderator

References

Enabling Interoperable Electronic Quality Measurement using HL7 FHIR®

Anne Smith, BSN, MHSA,1 Lisa Anderson, MSN, RN-BC,2 Rob Samples, MS,3 Moderator: Viet Nguyen, MD4

1National Committee for Quality Assurance, Washington, DC; 2The Joint Commission, Chicago, IL; 3ESAC, Incorporated, Rockville, MD; 4Stratametrics, LLC, Salt Lake City, UT

Abstract

Challenges to some of the Centers for Medicare & Medicaid Services (CMS) quality incentive programs have cited provider burden and feasibility concerns related to electronic clinical quality measures (eCQMs.) The move to use Health Level Seven International’s (HL7’s) Clinical Quality Language (CQL) for 2019 reporting allowed greater expressivity and more precise, unambiguous logic statements. However, retrieving data expressed in CQL is challenged by mapping requirements to access data entered and shared as part of healthcare delivery. This fact also limits sharing expressions with clinical decision support (CDS) artifacts to enable real-time workflow enhancements to improve performance. Panelists will describe how HL7’s Fast Healthcare Interoperability Resources (FHIR) and emerging HL7 implementation guides can improve measure precision and expressivity and provide insights into expectations for more straightforward implementation. Panelists will use examples of eCQMs developed using these new standards and tested in FHIR Connectathons and in real world pilots.

Learning Objectives:

- Understand the value of FHIR quality measurement on clinical systems including the potential for clinical burden reduction.
- Describe how FHIR quality measurement can mitigate the current challenges of eCQM management.
- Learn how new FHIR implementation guides can be used to represent clinical concepts needed for eCQM development, data retrieval, reporting.

General Description of the panel and the issue(s) that will be examined and a brief description of each panelist’s presentation.

The panel will discuss the current processes for electronic clinical quality reporting, evolution of eCQM standards, and associated challenges with quality measurement. eCQMs use data electronically extracted from EHRs to measure the quality of healthcare provided in order to identify opportunities for clinical improvement. The results of quality measurement are used throughout the healthcare system, including clinicians, researchers, and payers. Evolution of the existing standards to use CQL, a FHIR based data model, and FHIR Clinical Reasoning resources is necessary to align with rapid adoption of FHIR by the health information technology (IT) community as a platform for interoperable exchange of clinical information.

An eCQM is composed of three parts: the structure, the expression logic, and the data model. Figure 1 shows these measure components. The structure provides the comprehensive composition of a measure, including metadata (information about the measure) and the measure’s sections (e.g., population, numerator, denominator). It is not intended for human readability or consumption, but it can be represented using a style sheet for human readability.

The second part of a measure, the expression logic, enables articulation of information needed to determine if the subject of the measure meets the criteria for each of the measure’s sections. The expression addresses specific metadata about all data elements and their interrelationships required to meet such criteria. The third part of the eCQM, the data model, describes each data element and what metadata the expression logic can use to represent the measure’s intent.

Currently published eCQMs use a conceptual data model, the CMS Quality Data Model (QDM), to represent data, CQL as the expression logic, and Health Quality Measure Format (HQMF) as the structure. 2019 is the first reporting year using CQL as the expression logic.

![Figure 1. eCQM Structure.](image-url)
The data model represents the clinical concepts needed to measure and report quality. The QDM is conceptual, based on requirements of clinical quality measures used for manual data abstraction from clinical records. In the ten years since QDM’s creation, it has evolved to be more consistent with data that may exist in clinical settings. However, the data model is unique to the eCQM development environment and not implemented directly within a clinical environment. HL7 FHIR includes a set of resources based on data required for direct patient care, intended for defining semantically interoperable messages to communicate clinical, administrative, and financial information such that sender and receiver understand each other. FHIR has the potential to alleviate the additional burden of data model implementation by aligning eCQMs with other health data exchange scenarios.

Based on requirements defined in QDM and those defined in a CDS data model, and the virtual medical record (vMR), HL7 published a common approach based on FHIR for measurement and for CDS. The implementation guide, Quality Improvement Core (QICore) includes additional constraints on using FHIR for eCQMs and CDS artifacts to limit requests for data that may be excessive and add provider burden. Quality Improvement Clinical Knowledge (QUICK) further presents the QICore content in a logical view to help eCQM and CDS developers choose reasonable data within their CQL logic expressions. The current QICore implementation guide also provides a detailed mapping from QDM to guide eCQM developers considering transition to FHIR-based measures.

The FHIR Clinical Reasoning Module (Figure 2) incorporates the resources to express either an eCQM or a CDS artifact using the standard metadata, expression logic and data model. A measure report provides results, i.e., information retrieved about individual subjects measures (the individual report), evidence regarding which data elements were retrieved (based on the data model) and how the measured entity performed with respect to each measure section and the measure as a whole (aggregate report). Existing eCQMs use HL7’s Quality Reporting Document Architecture (QRDA). QRDA Category I is used for the single patient report and QRDA Category III for the aggregate report. FHIR Clinical Reasoning provides the Measure Report resource, based on QRDA, but interoperable with FHIR standards. It contains the patient, patient list, and aggregate reports that map to QRDA Category I and III, respectively; it also addresses reporting requirements that QRDA never adopted.

Currently system developers are required to use the HQMF, a standard based on HL7 V3 messaging to extract eCQM data from a system. When reporting quality data to a payer, this data is typically sent using QRDA, which is based on HL7 Clinical Document Architecture (CDA). This creates a misalignment between the standards for retrieving information from a system and producing a report for sending that information. FHIR Clinical Reasoning aligns these standards under a common framework.

The methodology used for generating, sending, and receiving quality reports can be a burdensome process and is prone to error. By aligning standards throughout the eCQM lifecycle, the panel will show how to improve the overall process and quality of data, while reducing the steps necessary for implementation. The use of Smart on FHIR Application Programming Interfaces (APIs) can provide a secure mechanism for exchange of data in a quality
reporting scenario. In addition, implementation of FHIR quality measurement has the potential to produce eCQM results in near real-time. Figure 3 shows the information flow for FHIR quality reporting.

An explanation why the topic of this panel is timely, urgent, needed, or attention grabbing is required with a discussion of anticipated audience.

Quality measurement programs continue to be an important topic that have the ability to impact clinical practices. This will continue to be the case as we transition to value-based healthcare at both the federal level and through private payers. As the requirements for reporting quality data increases, new methods are needed to promote interoperable exchange that align with industry standards and reduce the burden to care providers. FHIR is a new standard which is being rapidly adopted by the health IT community as a platform for interoperability. Nearly every EHR vendor and healthcare system in the US is involved in adoption of FHIR to some extent. If quality measurement is not included in this adoption cycle there is potential for misalignment with payment programs and systems development. In order to transition to FHIR quality measurement in a timely fashion, with minimal impact to providers, user feedback is needed. Audience participation for this panel is focused on getting this feedback to help shape implementation decisions and timelines.

Viet Nguyen, MD (moderator) will introduce the panelists and provide an overview of quality incentive programs and the need for improved quality measurement to enable value-based care. He will discuss FHIR’s potential to reduce the burden of reporting eCQMs by aligning with other data exchange scenarios, and discuss challenges with quality measurement in a provider setting. Dr. Nguyen will moderate the discussion.

Robert Samples will discuss the existing HL7 FHIR standards available to author eCQMs: the FHIR Clinical Reasoning Module, FHIR Quality Measure Implementation Guide, Data Exchange for Quality Measures (DEQM) Implementation Guide and Quality Improvement Clinical Knowledge as a logical data model and its alignment with CDS.

Anne Marie Smith will provide an example of an ambulatory care measure, comparing the existing version using HL7’s CQL-based HQMF with QDM to FHIR Quality Measure and DEQM Implementation Guides with the QUICK data model.

Lisa Anderson will provide an example of hospital-based care measure, comparing the existing version using HL7’s CQL-based HQMF with QDM to FHIR Quality Measure and DEQM Implementation Guides with the QUICK data model.

Discussion questions to enhance audience participation.

- How will transitioning to FHIR address the following challenges of quality measurement?
  - Access to real time information for measure performance
  - Data collection at the point of care
  - Accessing data needed for reporting from clinical systems
  - Exchanging quality data with vendors, payers, etc.
- How will aligning quality reporting with other data exchange scenarios impact clinical care?
- How would having a system that allowed a receiver of quality information to pull information without intervention be beneficial?
- What are the major challenges, roadblocks, or considerations of transitioning to FHIR for quality reporting?

All participants have agreed to take part in the panel.
Clinical text mining in mental health

Jessica D. Tenenbaum, PhD¹, Ramakanth Kavuluru, PhD², Thomas McCoy, MD³, Ozlem Uzuner, PhD⁴, Sumithra Velupillai, PhD⁵

¹Duke University, Durham, NC; ²University of Kentucky, Lexington, KY; ³Massachusetts General Hospital, Boston, MA; ⁴George Mason University, Fairfax, VA; ⁵King’s College, London, UK

Abstract

It has been shown that application of natural language processing (NLP) and text mining of clinical notes in electronic health records (EHRs) can help facilitate deeper and more accurate phenotyping in many disease areas. These methods are particularly useful in the context of mental illness, where disorders are highly heterogeneous and lacking diagnostic lab tests, medications are non-specific, and diagnosis codes are used inconsistently by different care providers, if not left out altogether. However, NLP approaches in mental health EHRs also pose a number of challenges, some common across all of medicine, some unique to mental illness.

The assembled panelists represent a geographically and scientifically diverse set of perspectives in the field of clinical NLP for mental health. They will discuss how their respective groups have addressed both scientific and infrastructural challenges and how they have helped to advance the field. They will also describe lessons learned as well as tools and artifacts they have developed that will be of use to other groups who are interested in similar questions.

Introduction

Mining electronic health record (EHR) data in the area of mental health is particularly challenging for a number of reasons. Psychiatric disorders are heterogeneous and lack diagnostic tests and robust biomarkers. Psychiatric medications are non-specific and the categorical diagnostic system is unreliable and used idiosyncratically. In light of this, natural language processing (NLP) is particularly powerful when applied to clinical documentation in the domain of mental illness.

Approaches to NLP span a continuum from simple bag-of-words to sophisticated grammar-based parse trees, with many levels of complexity in between. More recently, deep neural networks have demonstrated utility in analysis of free text. NLP may be used for cohort selection, deep phenotyping, identifying novel terms, enabling clinical decision support, and health outcomes research, among other purposes. Challenges in NLP include term extraction based on limitations of existing terminologies, the labor intensive nature of annotating datasets for training and testing, and the subjective and often ambiguous nature of clinical notes.

Topic importance and timeliness

Applying NLP methods on mental health EHR data holds particular promise not only because this type of documentation is predominantly written in free-text, but also because EHR data provides means to gather large samples for complex and rare clinical constructs that are otherwise challenging to pool. NLP in mental health has grown to be such an active field that it now has multiple satellite meetings devoted to it.¹³ Moreover, as neuropsychiatric illness is the leading cause of disability adjusted life years lost in the United States these illness are increasingly well resourced as clinical and scientific priority areas.⁴ ⁵

Speakers:

Dr. Jessica Tenenbaum- Moderator

Dr. Tenenbaum is an Assistant Professor of Translational Biomedical Informatics at Duke University. Dr. Tenenbaum is co-founder and Chair of AMIA’s Mental Health Informatics Working Group and Member-at-Large for the ELSI (Ethical, Legal, and Social Issues) WG. Current research projects focus on using EHR data for patient stratification in mental illness. Dr. Tenenbaum uses NLP to extract symptom-related terms from clinical notes for a machine learning approach to analysis and stratification beyond heterogeneous DSM-based diagnoses. Dr. Tenenbaum will introduce the panel and moderate the discussion.

Dr. Ozlem Uzuner- CEGS N-GRID Challenge, a.k.a., RDoC for Psychiatry Challenge
Dr. Uzuner is an Associate Professor at the Information Sciences and Technology Department of George Mason University. She also holds a visiting associate professor position at Harvard Medical School and is a research affiliate at the Computer Science and Artificial Intelligence Laboratory of MIT. Dr. Uzuner specializes in Natural Language Processing and its applications to real-world problems, including healthcare. Her current research interests include information extraction from fragmented and ungrammatical narratives for capturing meaning, studies of consumer generated text such as social media and electronic petitions, and semantic representation development for phenotype evaluation notes.

Dr. Uzuner will describe the CEGS (Centers of Excellence in Genomic Science) N-GRID (Neuropsychiatric Genome-Scale and RDoC (Research Domain Criteria) Individualized Domains) challenge to determine lifetime maximum symptom severity of patients in positive valence domain, based on their psychiatric intake interview reports. In this work, highlighted in Dr. Jim Cimino’s 2018 Year in Review talk at AMIA’s 2018 Symposium, 1000 reports were de-identified and annotated by experts for positive valence symptom severity.6, 7 The reports were shared with the community for de-identification and for symptom severity prediction of positive valence domain disorders. Findings from the shared task showed that a variety of approaches may succeed in extracting symptom severity information from these records. The shared task also demonstrated that mental health records differ from other clinical records in both content and presentation such that de-identification systems developed on discharge summaries may not automatically generalize to mental health reports.

Dr. Thomas McCoy- Annotation for gold standards

Dr. McCoy is the Director of Research at the MGH Center for Quantitative Health and Harvard Medical School Assistant Professor of Medicine and Psychiatry. His research focuses on development of computed phenotypes in, and applicable to, secondary use of healthcare data generated as part of routing care. His current interest is in developing methods to identify free text based phenotypes from expert knowledge without strict dependence on expert annotation.

Dr. McCoy will describe the painful practicalities of getting expert annotation of emerging transdiagnostic concepts like those in the CEGS N-GRID RDoC challenge and his team’s search for annotation-free methods to make these estimates.8 He will further describe how these methods were also conducive to scale out in the Partners Healthcare biobank where early genetic associations were identified9 and the full health record for use in suicide risk stratification.10 By removing the need for expert annotations this work was able to scale to cover the all domains of the NIMH RDoC system and by prioritizing run time simplicity it was able to scale up to both genetic discovery and clinical significant risk stratification within the confines of existing infrastructure.

Dr. Ramakanth Kavuluru- Deep Learning for NLP

Dr. Kavuluru is an Associate Professor in the Division of Biomedical Informatics (Internal Medicine) at the University of Kentucky. He primarily works in natural language processing for biomedicine using scientific literature, clinical narratives, and social media posts with the broad goal of extracting actionable information from these unstructured sources. High-level applications of his research include cohort selection for clinical trials, literature based knowledge discovery, knowledge-based exploratory search, and clinical decision support for precision medicine. He employs machine learning methods (including deep learning) and other data mining approaches to drive his research agenda.

Dr. Kavuluru will present two projects involving deep neural networks in clinical NLP, addressing how deep models compare with linear models for quantifying mental illness from clinical text. The first involves an overall ordinal four-level symptom severity scale as proposed by the 2016 CEGS N-GRID RDoC challenge.11 The second predicts specific mental health conditions based on the “history of present illness” portion of psychiatric evaluation notes.12 Results showed that linear models fall behind deep models in general but tend to complement deep models when building superior ensembles. Neural attention mechanisms also provide an excellent framework for error analyses (model debugging) and interpreting model predictions.

Dr. Sumithra Velupillai- NLP and Mental health outcomes research in a UK population

Dr. Velupillai is a Lecturer in Applied Health Informatics, at the NIHR South London and Maudsley Biomedical Research Centre (SLaM BRC), Institute of Psychiatry, Psychology and Neuroscience, King’s College London, UK. Dr. Velupillai has worked on clinical NLP since 2007 and on mental health-related research since 2016, particularly
on information extraction techniques. Dr. Velupillai’s team focuses on developing approaches for extracting temporal information as well as data-driven approaches for deep phenotyping and vocabulary development.

The Clinical Record Interactive Search (CRIS) system at the SLaM BRC provides a rich resource for NLP-based translational research. Knowledge-based, symbolic NLP approaches have been developed and applied to CRIS over the past 10 years to derive variables such as symptoms, medications, and risk factors from mental health records for subsequent large-scale clinical research. Extracting clinically important constructs from mental health notes using NLP poses some particular challenges, especially when NLP outputs are used for health outcomes research. Novel approaches are required for evaluation practices in this context. Suggestions for rigorous evaluation in mental health outcomes research will also be presented.

Questions to be addressed:

- What large-scale datasets exist for performing NLP in mental health?
- What are the biggest challenges for using clinical notes for deep phenotyping of mental health constructs?
- How is mental health different from, and similar to, the rest of medicine with respect to the opportunities afforded by NLP approaches? With respect to the challenges posed by use of NLP?
- What are some research purposes for which EHR data is “fit for use”, and what are some questions that will simply never be answerable by EHR data based on systematic bias, data missingness, etc?

Conclusion

This collection of panelists represents a diverse set of perspectives and approaches to the exciting and important topic of NLP in mental health. Panel attendees will learn about valuable community resources and cutting edge methods in this area and will be able to participate in discussion with panelists, adding their own experience and perspectives to the conversation.

References

Harnessing Data from mHealth Apps: Opportunities and Challenges for Clinicians and Researchers

Victoria L Tiase, MSN, RN-BC, FAMIA1; Robin R Austin, PhD, DNP, DC, RN-BC2; Christie L Martin, MN, MPH, RN, PHN, L_HIT-HP3; Ruth Masterson Creber, PhD, RN4; Spryros Kitsiou, PhD5

1Department of Information Technology, NewYork-Presbyterian Hospital, New York, NY; 2College of Nursing, University of Utah, Salt Lake City, UT; 3School of Nursing, University of Minnesota, Minneapolis, MN; 4Weill Cornell Medical College, Cornell University, New York, NY; 5University of Illinois at Chicago, Chicago, IL

Abstract
With the proliferation of mobile health (mHealth) technologies and applications (apps), patients are generating more data than ever before. These electronic patient-generated health data (PGHD) have the potential to yield additional insights, decrease nursing documentation burden, and improve patient care. At the same time, PGHD are also raising questions and posing significant challenges for researchers and clinicians. This interactive panel of nursing and consumer health informatics leaders will discuss key findings and challenges from several studies aimed at engaging patients with mHealth applications with the intent of collecting PGHD for research and clinical use, suggest a roadmap for additional research, and discuss future directions for nursing and informatics involvement in the design and implementation of mHealth apps.

General Description
According to the Pew Research Center, 77% of Americans now own a smartphone of some kind, meaning the majority of patients have powerful technology at their fingertips.1 The number of mHealth apps available for use has almost reached 300,000, making mHealth app development one of the fastest growing industries. In a recent report by Accenture, the majority (78%) of healthcare consumers wear or are willing to wear technology to collect data.2 These statistics signify a potential tsunami of valuable data about the patient condition that is not routinely captured from patients during initial assessments due to lack of time or poor recall. These data - also called the “digital phenotype” - can be used along with other information to broaden the biomedical profile of each patient and improve our ability to detect diseases or monitor their progression. The Office of the National Coordinator uses the term patient-generated health data to define these health-related data that are created, recorded, or gathered by or from patients to help address a health concern. PGHD, collected with mHealth apps, can fill in the gaps by providing important information about patients’ conditions between medical visits and on an ongoing basis. PGHD have the potential to efficiently capture needed information for use during care; contribute to cost savings; and improve the quality of care coordination and patient safety.

Unfortunately, there is very little evidence for healthcare professionals to base recommendations for the use of mHealth apps by clinicians or patients. Additional concerns surrounding the use of PGHD in clinical practice include time needed to review data, scope of practice, liability of the data received, and appropriateness of the interpretation and clinical decisions made based on the data.

As the most trusted healthcare professionals, nurses are on the receiving end of patient and family questions regarding which healthcare apps to download and where patients can provide the data that they have collected. Previous studies reveal that nurses are most often the clinicians interacting with patients about PGHD and mHealth apps.5 This puts nursing informatics researchers in a unique position to support work in this area. In order to effectively use PGHD, it is imperative to involve nurses in the design, implementation, and evaluation of mHealth apps and interventions.

The proposed interactive panel will engage a wide variety of clinicians, healthcare administrators and informaticians on a range of topics based on the collective work of the panel participants. Panelists will discuss challenges from several studies aimed at engaging patients with mHealth applications with the
intent of collecting data for research and clinical use, suggest a road map for additional research, and possible solutions associated with the development of mHealth interventions as well as the incorporation of patient-generated health data in electronic medical records for clinical decision making

Discussion Topics
Panel members will present their recent findings, personal perspectives, institutional experience, and engage audience members in a robust discussion to elicit participants’ opinions, experiences, and practices with mHealth applications. The following topics will be addressed:

- Assessing the need to engage patients and families with mHealth apps
- Preferred functionality of mHealth apps
- Implications of performing informatics research using data associated with mHealth apps
- Incorporation of mHealth data into electronic medical records
- Readiness and training of nursing staff in supporting the use of mHealth apps
- Role of nursing in the design and implementation of mHealth technology
- Security and compliance concerns related to mHealth data
- Impact of mHealth data on the nurse/patient relationship
- Nursing involvement in mHealth research initiatives

Moderator:
Spyros Kitsiou, PhD is an Assistant Professor in the Department of Biomedical and Health Information Sciences at the University of Illinois at Chicago and Director of the mHealth Innovation Lab. His research focuses on the development and evaluation of technology-based (mobile health and telehealth) interventions to promote healthy lifestyle behaviors and self-management of chronic disease (e.g. heart failure, hypertension, diabetes, asthma, and chronic obstructive pulmonary disease). His overarching research goal is to make a difference in people’s lives through the development of novel interventions, and evidence synthesis reports that target health outcomes and problems of significance to communities, governments, and health agencies. Dr. Kitsiou is currently conducting several randomized controlled trials of mHealth interventions with funding from NHLBI, NINR, and NIA.

Panel Members:
Victoria L Tiase, MSN, RN-BC, FAMIA is the Director of Research Science at New York-Presbyterian Hospital where she has been responsible for planning, organizing, and implementing a range of clinical information technology projects related to electronic health record adoption and meaningful use, clinical workflows, patient engagement, and care coordination. Ms. Tiase is a board-certified registered nurse with experience in perioperative nursing and nursing education. She is currently pursuing a PhD in Nursing from the University of Utah and was recently named a Jonas Nurse Leaders Scholar. Her research is focused on incorporating patient generated data into clinical workflows to impact care and reduce documentation burden. In 2014, she was selected for the Alliance for Nursing Informatics Emerging Leader fellowship. In addition, she currently serves on the AMIA Nursing Informatics Working Group Leadership team and is the Member-at-Large on the Alliance for Nursing Informatics Steering Committee. Ms. Tiase will present her work related to two patient-facing mHealth app pilot and the lessons learned with incorporated these data into clinical care. She will also discuss the policies and procedures of a large academic medical center in regards to mHealth apps.

Robin R Austin, PhD, DNP, DC, RN-BC is a Clinical Associate Professor at the University of Minnesota School of Nursing. Dr. Austin has a broad range of healthcare experience as a board-certified
registered nurse with a Master of Science in nursing focusing on nursing education and a Doctorate in Nursing Practice specializing in Informatics. She recently completed a PhD in Nursing at the University of Minnesota, School of Nursing where her research focuses on leveraging informatics for an integrated health approach to achieve optimal health for patient engagement, lifestyle modification for chronic disease, and health promotion and wellness using digital apps and mobile platforms. In addition, she is the current outgoing chair for the AMIA Consumer and Pervasive Health Informatics Working Group. Dr. Austin will discuss the literacy level of mHealth apps and her PhD research using structured consumer-generated mHealth data.

Christie Martin, MN, MPH, RN, PHN, LHT-HP is a medical-surgical nurse at Abbott Northwestern hospital in Minneapolis. She completed a Master of Nursing and a certificate in Leadership in Health Information Technology for Health Professionals from the University of Minnesota School of Nursing and a Master of Public Health from the University of Minnesota School Of Public Health. Ms. Martin’s research interests include child and adolescent health promotion and consumer health informatics. Ms. Martin serves as the co-chair of the Nursing Knowledge Big Data Mobile Health for Nursing working group, is an active member of the American Medical Informatics Association Consumer and Pervasive Health Informatics Working Group, and is the student representative for the University of Minnesota School of Nursing Research Committee. Ms. Martin will speak to the types of mHealth research that are being conducted by nursing informaticians and opportunities for collaboration.

Ruth Masterson Creber, PhD, RN is an Assistant Professor of Healthcare Policy & Research, Division of Health Informatics at Weill Cornell Medicine. Her research focuses on using mobile health to support patients with chronic disease management from diagnosis through the end of life and is funded by the National Institute of Nursing Research, R00NR016275 grant, “mHealth for Heart Failure Symptom Monitoring.” Throughout this work she has applied participatory design principles to the development of a web-app for older, chronically-ill adults with limited English proficiency, low health literacy and limited technology experience. She is also the Principal Investigator of a study focused on optimizing symptom management at the end of life for patients on hospice with the Visiting Nurse Service of New York. Dr. Masterson Creber will speak on the preferred functionality of mHealth apps, the role of nursing and healthcare informaticians in the design and implementation of mHealth technology and mHealth research initiatives.

Acknowledgements
Ms. Tiase would like to acknowledge the support of Jonas Philanthropies. Dr. Masterson Creber’s work has been funded by the National Institute of Health/National Institute of Nursing Research (R00NR016275).

References
5. Reading MJ, Merrill JA. Converging and diverging needs between patients and providers who are collecting and using patient-generated health data: an integrative review. Journal of the American Medical Informatics Association. 2018.
Informatics Architecture for the Future of Cancer Immunotherapy Research

Theresa L. Walunas, PhD¹, Abel N. Kho, MD, MS¹, Nike Beaubier, MD², Nikesh Kotecha, PhD²
¹Center for Health Information Partnerships, Northwestern University, Chicago, IL, USA, ²Tempus, Chicago, IL, USA, ³Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA

Abstract

Immunotheapy represents the next generation of cancer therapeutics and have shown dramatic impact and durable responses on a wide variety of tumor types. However, these therapies only provide durable outcomes for 20-40% of patients and up to 90% of patients are expected to experience a so-called immune related adverse event. Large scale EHR data from multiple institutions, linked to genomic information as well as other environmental data could be used as an underlying architecture to drive the future of precision medicine in cancer immunotherapy research. Participants in this panel discussion will learn more about 1) current lessons learned and challenges faced by those using informatics approaches to understanding cancer immunotherapy outcomes, 2) the availability of data to support precision theapeutics in cancer, 3) the challenges with using EHR data as a source for phenotypic and treatment data linked to genetic data to identify markers for successful treatment as well as risk factors for adverse events.

Introduction

Immunotheapy, including immune checkpoint inhibitors (ICI) and chimeric antigen receptor T cell (CAR-T) therapy, represent the next generation of cancer therapeutics and have shown dramatic impact and durable responses on a wide variety of tumor types. They are already in use in standard of care settings for melanoma, lung cancer and most recently breast cancer. However, these therapies only provide durable outcomes for 20-40% of patients¹ and 75-90% of patients² are expected to experience a so-called immune related adverse event (irAE) that range from mild rashes, to irreversible autoimmune complications, to death as a result of complications from severe inflammatory reactions. Understanding factors that impact treatment outcomes and the development of immune toxicity is critical for precision strategies for matching cancer patients with the therapies most likely to have the most significant impacts while minimizing adverse events.

While many clinical trials examining the function and outcomes of ICI and CAR-T therapies have been performed, they often represent small populations of people with cancer and outcomes and adverse events are tracked and recorded variably across studies, making it difficult to compare studies and develop generalizable conclusions³. In addition, these studies represent many different cancer types and may not represent a diverse patient population. Thus it is challenging to use these studies to develop effective models to understand risk factors that are associated with poor outcomes and the experience of adverse events. In fact, while there have been a few reports of biomarkers for successful therapy⁴ there are still relatively few means of predicting durable outcomes or the likelihood of experiencing adverse events and almost no studies that have demonstrated genetic linkages to outcomes.

ICI therapies have been approved for treatment of a variety of human cancers since 2011 and these therapies have expanded beyond clinical trials into common use. In 2010 the US Federal government enacted a series of reforms that spurred the adoption and use of electronic health record systems (EHR) across the US healthcare system. As of today, greater than 90% of hospitals and more than 80% of outpatient healthcare providers use EHRs to track patient care, including documentation of diseases and clinical conditions, treatments and laboratory test results. Providing a unique foundation on which to understand cancer therapy as well as to identify factors that improve or detract from therapeutic success. In addition, consortia, such as the NHGRI funded Electronic Health Records and Genomics network (eMERGE)⁵, have demonstrated that EHR data collected as a part of routine clinical care can be used to accurately phenotype patients across multiple institutions for well powered genomic analyses. Put together, this suggests a mechanism by which large scale EHR data from multiple institutions, linked to genomic information as well as other environmental data (such as a patient’s geolocation) could be used as an underlying architecture to drive the future of precision medicine in cancer immunotherapy research.
Panelists and Topics

1) Nikesh Kotecha, PhD is the Vice President of Informatics at the Parker Institute for Cancer Immunotherapy. Dr. Kotecha has expertise in computational immunology and biomedical informatics. He will moderate the discussion and will introduce the “The Informatics of ImmunoOncology – From Synthetic Biology to Translational Approaches Incorporating Real World Evidence for Managing Toxicities, a View from the Parker Institute” with a focus on the role of informatics in immuno-oncology using experiences from the Parker Institute, a new effort that has brought together pioneers in the immunotherapy space and 7 top cancer centers (Stanford University, UCLA, UCSF, M.D. Anderson Cancer Center, the University of Pennsylvania, Dana Farber Cancer Institute and Memorial Sloan Kettering Cancer Center).

2) Theresa Walunas, PhD is an Assistant Professor of Medicine and Associate Director for the Center for Health Information Partnerships at Northwestern University. Dr. Walunas has expertise in immunology and medical informatics and participated in early research studies elucidating the function of the first checkpoint inhibitor to be identified, CTLA-4. She will discuss “Managing Toxicities in Cancer Immunotherapy: Closing the Gap Between the Bench and Bedside to Improve Outcomes” with a focus on the intersection of clinical and research data to develop actionable, testable hypotheses to help improve therapeutic outcomes and reduce immune related adverse events.

3) Abel Kho, MD, MS, FACMI is an Associate Professor of Medicine and Director of the Center for Health Information Partnerships at Northwestern University. Dr. Kho has expertise in integrating EHR data across diverse care settings as part of several national networks including PCORnet 2.0, the All of Us Research Program, and EvidenceNOW. He will discuss “Integrating Medical Record Data as a Foundation for Cancer Immunotherapy Research – Challenges and Opportunities” and outline practical and efficient strategies for developing data pipelines that address data quality and granularity issues inherent in using EHR data.

4) Nike Beaubier, MD is Director of Pathology and Medical Lab Operations at Tempus. Dr. Beaubier has expertise in molecular genetic pathology with additional sub-specialization in gastrointestinal and surgical pathology. She oversees the laboratory operations at Tempus which processes over thousands of specimens per month. Dr. Beaubier will discuss "Building a High Throughput, Rapid turnaround, and High Quality Molecular Genetic Laboratory” with a focus on novel processes and tools developed to manage specimens, annotation, and analysis at scale.

Panel Rationale

Immunotherapy has advanced beyond clinical trials to become a core, mainstream cancer therapy that is now used in the treatment of a wide variety of tumor types. Given the wide reach of these therapies beyond the academic medical center and into community settings, there will be a dramatic expansion of real world clinical evidence that, when effectively linked to other data sources will be critical tools to understand factors that are predictive of durable outcomes in response to treatment as well as risk factors for severe immune adverse events.

Real world data sources come with a host of challenges that must be navigated successfully in order to create an effective platform for exploration of cancer immunotherapy outcomes and toxicities. This panel will explore lessons learned and the current state of the art in immuno-oncology informatics incorporating real world data sources; how real world data can be intersected with more traditional research data sources to create in silico disease models; the viability of EHR data as a foundation for large scale investigations and the intersection of EHR data with other key information (such as genetic information, cellular profiling and social and environmental determinants) to explore markers for successful treatment as well as risk factors for adverse events.

This panel is a multidisciplinary team reflecting complementary perspectives on the science and technology required to develop effective informatics architecture for exploring immuno-oncology. Drs. Kotecha and Walunas will be able to reflect on key research questions and challenges in the field and Drs. Kotecha, Kho will frame the data space, informatics infrastructure and challenges and Drs. Kho and Beubier will discuss the intersection of genomics and pathology data with real world clinical data.

The anticipated audience for this presentation would be those interested in understanding the current challenges facing cancer immunotherapy informatics, current data science approaches for exploring cancer immunotherapy, and those who are interested in developing the next generation infrastructure for informatics-based cancer immunotherapy research and care quality improvement.
Discussion Questions

• How do we intersect clinical information with bench information and develop frameworks that can be used to develop actionable, testable insights?

• What kinds of tools would be essential additions to an integrated EHR-based data platform to support effective investigations and identification of biomarkers?

• In addition to the data, what kinds of people, i.e. data scientists, clinical researchers, clinicians and basic scientists are needed to provide the subject matter expertise to develop an effective platform for research?

• How can patient reported data be integrated into this architecture? How could it best be used? What would be the existing challenges to integrating patient reported and wearables data?

• What are the best strategies for integrating human data with non-human animal models to support mechanistic studies that cannot be performed in humans?

• How do we protect patient privacy in this system? Cancers are often rare and cancer + a specific therapy even rarer.

• What are the experiences from an end user perspective? What are the clinicians struggling with compared to the physician researchers? How are they manifesting themselves in different cancer specialties?

Organizer Statement

All of the participants in this panel have reviewed this abstract for this panel and have agreed to take part in the panel should it be accepted.

References

Interpretability and Statistical Inferences of the Prediction of Clinical Outcomes Using Deep Neural Networks

Qing Zeng-Treitler, PhD a,b, Yijun Shao PhD a,b, Cecilia Dao, PhD c,d, Joseph Goulet PhD c,d, Orna Intrator, Ph.D e,f

a Department of Biomedical Informatics Center, George Washington University, Washington DC, USA,
b Washington DC VA Medical Center, Washington DC, USA
c Department of Medical Informatics Center, Yale University, New Haven, Connecticut, USA
d West Haven VA Medical Center, West Haven, Connecticut, USA
e Department of Public Health Sciences, University of Rochester Medical Center, Rochester, New York, USA
f Canandaigua VA Medical Center, Canandaigua, New York, USA

Abstract
Deep neural network models (DNNs) are being explored in the clinical domain, following the recent success in other domains such as image recognition. For clinical adoption, outcome prediction models require explanation; however, due to the highly non-linear inner transformations, DNNs are viewed by many as a black box. In this panel, we will discuss the interpretability and statistical inferences of the prediction using DNN model, in the context of real clinical outcome prediction use cases. The panel will begin by providing a brief overview of the deep neural network and latest research to interpret DNN results. We will then describe two novel explanation methods. We will also introduce methods to construct confidence intervals and p-values for the explanations and validate them through simulations and data application. Finally, we will discuss the challenges and future work to improve the understanding and acceptance of DNN models by researchers and clinicians.

Learning Objectives
1. Understand the challenges and needs of explaining deep neural network models in the context of clinical research and clinical decision support.
2. Describe state-of-the-art approaches to explain deep neural network models.
2. Compare the explanations of deep neural network models with statistical regression models.

Introduction
Deep neural network (DNN) models are one of the key approaches in artificial intelligence (AI). Recent studies have shown that DNN can outperform traditional machine learning (ML) methods and statistical methods in clinical and non-clinical domains1-8. To influence clinical practice, a prediction model must also have sufficient discriminating power, calibration, face validity and clinical utility. Thus, to gain clinical acceptance and impact, we must provide interpretation of the DNN models. Traditional statistical models, especially linear models such as logistic regression, are popular not because they often have the best prediction performance, but rather they provide simple interpretations of each independent variable’s significance in the prediction process (e.g., in the form of odds ratio). However, the DNN models are particularly difficult to explain due to the highly complex transformations carried out between the “deep” layers.
DNN models are viewed as black boxes, and understanding and explaining their behaviors is the subject of ongoing research. Compared to clinical data, images of natural scenes, recordings of speech, or game board layouts are much easier for humans to grasp. As a result, most published work about the DNN explanation focus on confirming human interpretation, e.g., if an image feature critical to the prediction of being a dog looks like the outline of a dog. With clinical prediction, human users do not have such prior knowledge and expect the model to yield information that justifies the outcome prediction.

Approaches are being explored to explain DNN models include Attention Mechanism, Local Interpretable Model-Agnostic Explanations (LIME) and Layer-wise Relevance Propagation (LRP). Each of these approaches has strengths and limitations. One key challenge is to interpret and draw statistical inferences based on the explanations generated by these approaches. Another key challenge is to validate the DNN explanations. With background in biomedical informatics, computer science, mathematics, statistics, and outcomes research, our panel of experts will provide an overview of explainable AI methods and describe two explainable AI methods in more detail.

This panel is intended to address biomedical informatics professionals, who would like to learn about the new approaches in explainable AI and their application in specific clinical uses. This topic is timely as there is increasing use of deep neural networks in the biomedical domain while their adoption in the clinical settings is limited by the difficulty of explaining them.

Presentations

The first panelist, Dr. Zeng-Treitler, is an informatics expert with a background in computer science. Dr. Zeng-Treitler will provide a brief introduction to DNN and describe the need to explain DNN models in the context of several specific clinical research projects, including the prediction of post-surgery outcomes, prediction of outcomes in patients with serious mental illnesses, and the assessment of severity of illness in hospitalized pediatric patients. She will then compare and contrast several existing explainable AI approaches: Attention mechanism provides a window to observe how features are utilized by a DNN but its explanation is tied to the DNN design. LIME is agnostic about a model’s inner structure and assesses the role of a feature purely based on the change in the output. LRP starts with an input (i.e. a set of features) and the probability of an output. It works backwards layer by layer to redistribute this probability to each input feature. A relevance score is then computed for the feature.

The second panelist, Dr. Shao, is an informatics expert with a background in theoretical and applied mathematics. He will describe a novel LIME method called Impact Assessment to explain DNN. To measure the importance of each variable on the predicted outcome, the Impact Assessment method defined impact scores on observational, individual, and population levels. The Impact Assessment method was validated by comparing its results with odds ratios derived from logistic regression, showing strong but not perfect correlations as expected. Another validation approach is to use a dataset with known underlying relationship between predictors and outcome. Without complete knowledge of underlying relationships in real patient datasets, a simulated dataset with 100 variables and a mixture of linear and none-linear relationships was generated. Noting that simulated data are far less complex that real patient datasets, the Impact Assessment was shown to be highly accurate (90-95%) in estimating the contribution of a variable to the outcome.

Third panelist, Dr. Dao, expert in biostatistics with a background in statistics, will talk about the statistical inferences of the impact score. She will describe the idea behind the non-parametric bootstrap method for constructing confidence intervals without assumptions. Then, a parametric method for confidence intervals and p-values based on large sample theory can be used for faster computation, given the assumptions are reasonable. The theory and intuition behind these methods will be discussed and evaluated based on the simulations and data application.

Fourth panelist, Dr. Intrator, is a health services researcher trained in statistics and applied mathematics. She will introduce a method that demonstrates the effects of inputs on output of neural networks by using novel robustification techniques. Simulated data from known models are used to demonstrate the interpretability results of the neural networks. Graphical tools are used for studying the interpretation results, and for detecting interactions between covariates. The effects of different regularization methods on the robustness of the interpretation will be discussed. An application to an ANN model predicting 5-year mortality following breast cancer diagnosis will be presented.

Moderator, Dr. Goulet, is the Director of the Methodology and Biostatistics core of the VA’s Pain Research, Informatics, Multi-morbidities, and Education (PRIME) HSR&D Center of Innovation, and PI of VA funded administrative data-based studies including the Association of Complementary and Integrative Health Interventions with
Opioid Use and Related Risks among Veterans with Musculoskeletal Disorders and PTSD. He will lead the discussion on the interpretation of DNN results in the broader context of clinical research.

**Discussion Questions**

- What are the pros and cons of training traditional statistical models, traditional machine learning models, and deep neural networks in clinical research?
- What types of DNN explanations do we need to enable clinical adoption?
- How to balance accuracy and interpretability in the design of predictive models?
- What roles do ground truth and human validation play in the development of explanation methods?
- What is the meaning of “statistical significance” in the context of DNN explanation?
- What are the key challenges in the understanding of DNN models as well as future steps?

**Participation Statement**

I, Qing Zeng-Treatler, hereby confirm that all speakers have agreed to participate in this panel. Panelists are aware that there are no travel funds available.

**References**

Informatics approaches to collecting, analyzing, and addressing social determinants of health in healthcare

Yiye Zhang, PhD,1 Evan Sholle, MS,2 Marianne Sharko, MD, MS,1 Yongkang Zhang, PhD,1 Jessica S Ancker, MPH, PhD,1
1. Weill Cornell Medicine, New York, NY; 2. Information Technologies & Services Department, Weill Cornell Medicine, New York, NY

Abstract

It is well-established in the public health literature that health is strongly influenced by socioeconomic, cultural, and environmental conditions, community factors, and lifestyle. As healthcare organizations are becoming interested in these social determinants of health (SDH), informatics has multiple opportunities to shape and facilitate the collection, analysis, and application of SDH data to improve health equity. In this panel, we discuss multiple challenges and opportunities around three areas: how to collect SDH data; how to analyze these data; and how to apply these data to improving healthcare operations and health equity. Panelists will also analyze methodological reasons why, when the direct effects of SDH on health are so clear, modeling approaches using large clinical data sets sometimes fail to find effects of SDH. After participating in this session, attendees should be able to define SDH, identify multiple barriers to the efficient and accurate collection and analysis of these data, describe several informatics approaches to overcoming these barriers, and demonstrate how SDH influence the course of a chronic disease (heart failure).

Keywords

Social determinants of health, socioeconomic status, electronic health records, geocoding, natural language processing

Introduction

Social and environmental factors such as education, housing, employment, living and working conditions, social relationships and networks, and lifestyle are strong predictors of health outcomes. These factors have historically been considered beyond the purview of medicine. However, current and pending changes to payment and reimbursement models in the United States are encouraging healthcare organizations to reframe their domain more broadly and consider ways to identify and address the influences of social factors on medical outcomes.

This increased interest is manifested in multiple new projects nationwide to infer SDH from patients’ home residence or from other data, administer questionnaire screening to collect SDH directly, incorporate SDH into research data sets and data models such as i2b2 as well as directly into electronic health records, and explore how to adjust for SDH in reimbursement models and risk models.

In this panel discussion, we demonstrate how the informatics community can play a central role in helping to collect, analyze, and apply data on social determinants of health (SDH) to improve health equity and patient outcomes. Panelists will discuss the current landscape of SDH data collection, including existing tools such as electronic health record-based screening instruments. We also discuss some of the potential problems and unintended consequences of incorporating SDH into healthcare, and will invite the audience to join in discussions about improving the collection and utility of these variables.

Panel Moderator: Why Are SDH Important for Improving Health Equity?

Jessica S Ancker, MPH, PhD will open the panel with a short overview of public health and informatics research on the impact of social determinants of health globally, ongoing informatics innovations around SDH data nationwide, and the implications of leveraging SDH data for health equity. She will also discuss some of the potential harms and unintended consequences arising from SDH adjustment or collection.
Dr. Ancker, associate professor at Weill Cornell Medicine, has conducted extensive research on health literacy and numeracy as SDH and is also leading a retrospective cohort study that combines clinical data with SDH data to examine the impact of SDH on 30-day hospital readmission.

Combining Neighborhood with Clinical Data Shows That Parks Are Good for Your Health

Yiye Zhang, PhD, will describe how, by linking electronic health record data with publicly available data about patients’ place of residence, her team of health informatics and environmental engineer researchers demonstrated that living farther from parks in New York City increases the chance that heart failure would worsen. The relationship holds true even when controlling for demographics, clinical risk factors, and neighborhood poverty rates. This rich combined dataset of clinical indicators and neighborhood metrics also revealed other aspects of the built environment associated with disease progression. In this talk, she will discuss geographic mapping methods and the analyses linking individual-level and neighborhood-level social determinants data with heart failure outcomes.

Dr. Zhang conducts research on developing and applying methods for analyzing the progression of health and healthcare utilization for the purpose of clinical decision support. She is interested in understanding how non-clinical factors including SDH and the built environment may affect disease progression and access to health.

Extracting SDH From Medical Records Reveals Biases in Data Collection

Evan Sholle, MS, will show how, despite regulatory efforts to encourage structured documentation of race and ethnicity, many patients still lack informative data in the electronic health record. In a recent study, WCM informatics and health services researchers collaborated to use natural language processing information extraction techniques to infer race and ethnicity from free-text notes in the EHR. The results showed that a substantial portion of patients described as Black and/or Hispanic in notes were not documented as such in structured EHR fields. Of greater concern, patients with missing or inconsistent data differed from those that were recorded as Black and/or Hispanic in structured documentation in significant ways. In this talk, he will discuss the evidence these findings provide for systemic bias behind the capture of SDH documentation in the EHR, and the need for informatics methods to fill in these gaps.

Evan Sholle is the manager of research informatics services at Weill Cornell Medicine. His research centers on the secondary use of electronic patient data to support clinical and translational research.

Should We Be Asking Patients About Their SDH?

Marianne Sharko, MD, MS, will address issues surrounding collecting potentially relevant SDH that are known only to the patient. For example, literacy level, history of domestic violence, or food insecurity would presumably be most accurate if reported directly by the patients affected. Several large-scale efforts to screen patients are ongoing. Unfortunately, screening for SDH (in addition to being pragmatically challenging) also has the potential to alienate or alarm patients. Some initiatives have shown that patients feel embarrassment about physicians recording some of these variables, or worry that their answers may be used to harm them. In this presentation, Dr. Sharko will discuss challenges and opportunities surrounding the direct collection of SDH.

Dr. Sharko, a pediatrician who received her master’s degree in health informatics at Weill Cornell Medicine, has research interests including the study of the unique privacy and confidentiality concerns for vulnerable patient populations in the context of electronic health records, patient portals, mobile health technology, and health information exchange.

When SDH Do Not Show an Effect, What Are the Reasons Why?

Yongkang Zhang, PhD, will show how, despite the demonstrated strong connections between social factors and health outcomes, many recent literature and modeling projects demonstrate that when SDH are added to models that already contain clinical data, they make little difference to predictive accuracy. This talk will analyze some of the previously published literature for reasons why SDH failed to predict health-related outcomes, and present analyses from our own data sets comparing predictive models with and without SDH. We will discuss how data quality
limitations, correlations between causally related variables, and healthcare utilization patterns of socially vulnerable patients explain some of the situations in which predictive accuracy is not improved by adding SDH.

Dr. Zhang’s research focuses on improving care delivery to high-need, high-cost patients, including those with vulnerable social conditions. His current project examines the association between SDH and healthcare utilization among Medicare and dual-eligible patients.

**Learning Objectives:**

After participating in this session, attendees should be able to:
- define social determinants of health;
- identify multiple barriers to the efficient and accurate collection and analysis of these data;
- describe several informatics approaches to overcoming these barriers;
- demonstrate how SDH influence the course of a chronic disease (heart failure).

**Panel Organizer Statement**

All participants have agreed to participate in this panel.

**Conflict of Interest**

The panelists have no conflicts of interest to disclose. Dr. Ancker’s healthcare disparities research is funded by the Office of the Weill Cornell Dean (Dean’s Diversity and Healthcare Disparities Grant Award), the National Library of Medicine (R01 LM012964), and the Patient-Centered Outcomes Research Institute Health Systems PCORnet (Patient-Centered Clinical Research Network) Demonstration Project (HSD-1604-35187, PI: Kaushal). Dr. Zhang is also funded by the Office of the Weill Cornell Dean (Dean’s Diversity and Healthcare Disparities Grant Award) and the US Department of Transportation-funded Center for Transportation, Environment, and Community Health. Dr. Zhang is funded by the Patient-Centered Outcomes Research Institute Health Systems PCORnet Demonstration Project (HSD-1604-35187, PI: Kaushal).
Team-Centered Informatics: A Necessary Adaptation to Translational and Implementation Science?  
Suresh K. Bhavnani PhD1, Shyam Visweswaran MD PhD2, Erich Kummerfeld PhD3, Carlos Clark DO4, Rebekah Penton DNP RN5

1Inst. for Translational Sciences, 4Dept. of Int. Med., 5School of Nursing, Univ. of Texas Medical Branch, Galveston, TX; 2Dept. of BMI, Univ. of Pittsburgh, Pittsburgh, PA; 3Inst. for Health Informatics, Univ. of Minnesota, Minneapolis, MN

Introduction

The rapidly maturing fields of translational science1 and implementation science2 have recognized the multidimensional nature of health outcomes, and the need for a holistic and integrative approach to address them. For example, several NIH-funded organizations have assembled multidisciplinary translational teams (MTT)3 consisting of biologists, clinicians, informaticians, and statisticians to help accelerate the integration of multi-omics data across the basic-clinical-community continuum. Similarly, implementation scientists draw from many disciplines to overcome cultural, political, and technical hurdles that prevent the broad uptake of research results produced from translational science.

Unfortunately, such teams face numerous barriers in integrating multidisciplinary knowledge, ranging from differing conceptualizations of a problem,4 to narrow social and technical identification within a single discipline.5 These hurdles often result in many teams having “low integrative capacity”.6 One approach, proposed by team science researchers to increase integrative capacity in teams, is through the use of “boundary objects”,7,8 which are external representations of information that facilitate the integration of diverse disciplinary knowledge. For example, as shown in Fig. 1, a boundary object could be designed to enable members of a multidisciplinary translational team to integrate their disciplinary knowledge, with the goal of generating novel insights that transcend individual disciplines.9

However, while boundary objects are used by the team science community to improve integrative capacity in teams, current informatics tools are mostly targeted to individual researchers. For example, powerful tools such as CRISPR10 and Genome Browser11 enable deep analyses of a few biological datatypes, but provide little support to investigators for integrating and interpreting results across multiple tools and disciplines. We therefore pose the question: Could the fields of translational science and implementation science benefit by a new class of informatics tools that are team-centered?

Goals and Structure of the Panel

Building on the concept of boundary objects from team science, this panel will explore the concept of “team-centered informatics”9 designed to help members of multidisciplinary teams integrate knowledge from diverse disciplines. Fig. 2 shows an example of a computational boundary object, which simultaneously models both molecular and clinical information. Being computational, this representation can be partitioned, filtered, layered, and intersected with different combinations of variables from each discipline to enable comprehension of how they are related to each other. Such a representation on the one hand could be meaningful to individual experts specializing in one dimension of the data such as genes, but on the other hand could also provide an integrated understanding of how those genes interact with other dimensions such as the environment, leading to insights that transcend individual disciplines. However, to make such representations more amenable to teams, we need to develop new interaction methods and algorithms for enabling the real-time collaborative exploration of diverse information in massive datasets.

To debate whether we need to invest in such team-centered informatics tools, this panel brings together a diverse set of researchers from human-computer interaction, machine learning, causal modeling, hospital administration, and nursing. Each panelist will present a use case from their research to explore two challenge questions: (1) is team-centered informatics old wine in new bottles? In other words, do collaborative tools already exist that are sufficient for teams to integrate diverse disciplinary knowledge? (2) does team-centered informatics provide low bang for the buck? In other words, does the extra cost of building new ways to interact with complex data provide much added benefit to researchers?
Position Statements

1. **Why do we Need Team-Centered Informatics?** (Moderator: Suresh K. Bhavnani, PhD)

Dr. Bhavnani, is professor of biomedical informatics at the Institute of Translational Science in the University of Texas Medical Branch, with a background in human-computer interaction. As PI of the Discovery and Innovation through Visual Analytics lab, he specializes in (1) the discovery of complex patterns in big biomedical data such as discovering heterogeneity in diseases and their respective pathways,12 and (2) the innovation of novel visual analytical methods to analyze and comprehend large and complex biomedical datasets.13 His research has received 5 distinguished paper awards from AMIA, and an outstanding paper award from the Science of Team Science (SciTS) conference.

Dr. Bhavnani will provide a brief overview of the motivations underlying translational science, and implementation science, and how they both benefit from conducting research through multidisciplinary scientific teams. Next, he will discuss the importance of external representations of data and knowledge such as visualizations and flow diagrams for (1) reducing a researcher’s cognitive load and freeing up resources to enable problem solving,14 and (2) as a “boundary object”,7 to help other members in a team span disciplinary boundaries, leading to novel insights.9 Using examples from translational science and implementation science, he will argue that the concept of a boundary object from team science could be the foundation for team-centered informatics solutions.9 He will then ask the panelists to present use cases from their research, and address the two challenge questions: (1) *Old wine in new bottles?* (2) *Low bang for the buck?*

2. **What is the Role of Team-Centered Informatics in Clinical Decision-Making?** (Shyam Visweswaran, MD PhD)

Dr. Visweswaran is associate professor of biomedical informatics in the Department of Biomedical Informatics and the Intelligent Systems Program, at the University of Pittsburgh. He specializes in the application of artificial intelligence and machine learning to problems in clinical medicine and translational bioinformatics. As PI of the Vis lab, his research includes (1) computer-aided diagnosis and prediction, (2) discovery and prediction from high-dimensional genomic data, and (3) patient-specific predictive modeling for personalized medicine.

Dr. Visweswaran will present a use case from his research15 focused on the use of Bayesian modeling and decision trees for clinical decision-making in multidisciplinary clinical teams. He will then address the two challenge questions with a focus on explainability of machine learning models in clinical decision-making.

3. **What is the Role of Team-Centered Informatics in Identifying Causal Mechanisms?** (Erich Kummerfeld, PhD)

Dr. Kummerfeld is research assistant professor in the Institute for Health Informatics at the University of Minnesota. He specializes in logic and computation applied to developing novel algorithms for discovering causal relationships, including relationships that involve, are between, or are confounded by latent variables, and applying them to health data to inform new treatments.

Dr. Kummerfeld will present a use case from his research16 focused on the use of causal modeling in biomedical data, focused on the identification of disease mechanisms and design of targeted interventions. He will then address the two challenge questions with a focus on the social dynamics of collaborations between data scientists and domain experts.

4. **What is the Role of Team-Centered Informatics in Hospital Administration?** (Carlos Clark, DO)

Dr. Clark is Chief Medical Information Officer (CMIO) at the University of Texas Medical Branch hospital. He specializes in quality improvement including decreasing the smoking rate of clinic patients, improving the transition of care to reduce hospital readmission, and leads endeavors in population health for alignment of UTMB as a high value practicing organization.

Dr. Clark will present a use case from his research17 focused on enabling providers to practice evidence-based medicine through EMR decision-support. He will then address the two challenge questions with a focus on overcoming social, political, and technical hurdles in implementing evidence-based practices.

5. **What is the Role of Team-Centered Informatics in Designing and Implementing Discharge Protocols for Skilled Nursing Homes** (Rebekah Penton, DNP RN)

Dr. Penton is Assistant Professor of Nursing in the School of Nursing at the University of Texas Medical Branch. Her research focuses on using methods from implementation science, and her 10 years of professional experience as a nurse practitioner, to improve care for geriatric patients in skilled nursing homes (SNFs).

Dr. Penton will present a use case from her research focused on the design and implementation of interventions to reduce hospital readmissions from SNFs. She will then address the two challenge questions with a focus on overcoming implementation hurdles related to hospital-SNF communication.
Discussion and Engagement of Panel Attendees

After each of the above position statements, the attendees will be first asked (with a show of hands) how many agree or disagree with the answers provided to the two challenge questions. For example, if a panelist claims that team-centered informatics is “old wine in new bottles”, then the attendees will be asked if they agree or disagree with that conclusion. The moderator will then invite attendees to voice their opinions about the issues being discussed. The panel members will engage with these opinions, and the moderator will determine when the next position statement will begin. After all the positions statements are delivered and responded to, the attendees will be asked through a final show of hands to determine if team-centered informatics is “old wine in new bottles” or not, and whether it has “low bang for the buck or not”. The results will be summarized by the moderator, and if the discussions lead to compelling arguments, will be documented in a report for publication. The first author has used the proposed panel format at the AMIA Informatics Summits in 2015, which resulted in high engagement among the panelists and attendees.

We expect this debate to enable attendees to comprehend the issues related to the design and use of team-centered informatics tools for a wide range of team-centered tasks in translational science and implementation science. The panel discussion should be of benefit to (1) researchers engaging in translation science and in implementation science, and (2) program managers from funding agencies who wish to understand how best to support the development, evaluation, and use of team-centered informatics approaches that have the potential of wide impact in making medical discoveries, and implementing them to improve patient care.

Acknowledgements

This research is funded in part by UTMB CTSA grant (UL1 TR001439), and PCORI award (ME-1511-33194).

The first author affirms that all panel members have agreed to participate, and have contributed to the preparation of this document.

References

Thriving in Your Biomedical Informatics Career
While Balancing Work, Personal, and Family Life

Sponsored and presented by the Women in AMIA Career Advancement Committee (CAC)
Lesley Clack, MS, ScD1, Qing T. Zeng, PhD2, Gretchen P. Jackson, MD, PhD, FACS,
FACMI, FAMIA3, William Hersh, MD, FACP, FACMI4, April Mohanty, MPH, PhD5

1University of Georgia, Athens, GA, 2George Washington University School of Medicine and Health Sciences, Washington, DC; Veterans Affairs Medical Center (VAMC), Washington, DC, 3IBM Watson Health, Cambridge, MA; Vanderbilt University Medical Center, Nashville, TN, 4Oregon Health & Science University, Portland, OR, 5University of Utah School of Medicine, Salt Lake City, UT; Salt Lake City VAMC, Salt Lake City, UT,

Abstract
A career in biomedical informatics typically requires a significant time and energy commitment, which must be balanced with other areas of life. An informatics professional can serve in multiple roles, such as domestic partner, dependent caregiver, etc. which may be challenging to their career trajectory and opportunities and may be hard to blend, though all of these roles can contribute to an enriching and fulfilling life. Biomedical informatics is especially challenging, given its interdisciplinary nature and wide breadth and depth. Key topics relevant to work-life balance include identifying values and priorities, creating a career matrix/timetable, family planning (e.g. family leave, dependent care), time management, burnout and managing daily struggles, work arrangements, dual-careers. In this interactive panel, a diverse group of informatics professionals will present different perspectives and discuss relevant issues, resources and recommendations related to balancing personal and family life while advancing their biomedical informatics careers. As these issues affect women and men, all informatics professionals who are interested in gaining insight about and advancing their informatics careers from various perspectives and engaging in a discussion about work-life balance issues are encouraged to attend. This interactive panel supports AMIA’s strategic directive of professional growth and leadership development of our members.

Learning Objectives: 1) Describe the issues related to work-life balance in biomedical informatics raised by scientific literature and from various perspectives including clinical, non-clinical settings and domains such as government, industry, and academia and across a wide range of career stages. 2) Describe the resources and solutions available to help caregivers achieve success both at home and in biomedical informatics professions.

Introduction
Studies of academic and industry professionals suggest that family responsibilities can have a major impact on career advancement and contribute to the salary gap. For instance, one study suggests that family formation has the “largest negative effect during the early stages of a woman’s career and that each child depresses a woman’s salary by 5%.” Another study found that women in science with children are 35% less likely enter tenure-track positions than men who are married with children. There are major gaps in research into work-life balance issues faced by informatics professionals, issues faced by the myriad family and work arrangements that exist such as single parents, minorities, LGBTQ members, male caregivers, dual career couples, super-commuters. Also, there have been few formal discussions on resources associated with success in achieving work-life balance in biomedical informatics.

Aim of the Discussion: Our aim is to facilitate a positive discussion regarding the current issues faced by male and female biomedical informatics professionals with multiple and competing life roles and to share resources, insights, and recommendations about achieving a balanced personal and family life while advancing professionally. Our panel will cover a variety of perspectives from different career stages, work settings (i.e. academic, healthcare, government, and industry), and backgrounds and will present on research, research gaps, and personal experiences,
to facilitate an interactive discussion with the audience. Examples of discussion topics and questions, in addition to those previously mentioned, include:

1. **Work-life balance issues in STEM professions identified by the scientific literature and strategies to support balance from the employer and employee perspective.** (Qing Zeng) The number of publications discussing work-life balance is increasing. About 70% of men and women in the workforce report work-life conflict. Motherhood poses additional challenges. Social and employment contexts also play an important role in work-family balance. Difficulties in work-life balance can lead to burnout and affect physical and mental health. Commonly recommended strategies include priority setting, self-assessment, support network, time management, flexible workplace and work schedules, and segmentation of work and life. No work life study has focused on biomedical informatics specifically, but research has been done in related STEM fields.

2. **Planning for and Managing Family as a Clinician Scientist and Healthcare Tech Executive** (Gretchen Purcell Jackson) Many clinicians and clinician-scientists face prolonged job-related stress across many workplace settings and there are gender differences in time spent on parenting/domestic responsibilities. For example, among 1,049 junior physician-researchers, women (vs men) were more likely (86% vs 45%) to have spouses or domestic partners who were employed full-time. Topics to be covered include considerations for, and potential solutions or resources relevant to planning for family, parental leave, caregiver guilt, work-work balance (i.e. juggling clinical, research and education responsibilities), returning to work, transitioning from academia to industry and into leadership positions and pros/cons related to work life balance for various workplace settings, and strategies for staying sane, happy and well.

3. **Work-life balance issues from the perspective of a male leader, educator, and mentor.** (William Hersh) While there are many work-life balance issues that are unique to women, some are common to all professionals such as prioritizing relationships, hobbies/outside interests, health and wellness, and avoiding burnout. Also, more men are facing the work-life balance struggle, as suggested by the National Science Foundation 1977 and 2008 surveys of dual-career couples, where stress from work-family conflict increased, from 35% to 60% in men vs 41% to 47% in women. Also, more men are taking on more domestic responsibilities, are married to, or interact with professional women. Relevant discussion topics include: How are work-life balance challenges faced by men changing? What are the common issues faced across gender/caregiver-role? What challenges are unique to dual-career professionals and related strategies and resources? How do we enable and empower mentees to build sustainable work-life balance? What are leaders and institutions doing to empower women to develop and thrive? How do we engage everyone on these issues and avoid making this a “women’s issue”?

4. **In the Trenches- new mom-new academic faculty life hacks and resources for caregivers and minorities at early career stages.** (April Mohanty) It has been said that “a female faculty member who is interested in starting a family finds that her biological clock is directly at odds with her tenure clock.” Many new mothers (or caregivers) cope with their responsibilities by “satisficing”. Relevant questions to junior faculty for discussion include: What are some examples of “satisficing”? How can junior and minority faculty develop a professional and personal environment that supports their success in their various roles as caregiver, professional, and personal life and what are some available resources? What are some time/stress saving tips for new parents?

**Intended Audience** includes all biomedical informatics professionals, including both women and men, who would like to learn about balancing multiple life roles or are interested supporting colleagues with multiple roles. This topic is timely as there is limited research on work-life balance issues in the biomedical informatics field and a growing number of potential/current caregivers are entering the workforce.

**Lesley Clack, ScD, MS** is Assistant Professor and Member of the Health Informatics Institute at the University of Georgia College of Public Health in Athens, GA. She teaches undergraduate and graduate courses across the informatics spectrum. Dr. Clack is actively involved with professional informatics associations, serving as Member on the AHIMA Informatics and Data Analytics Practice council, Member of the University System of Georgia Health Informatics Advisory Board and member of the Women in AMIA CAC. Dr. Clack will moderate the panel.

**Qing Zeng, PhD** is Director of the Biomedical Informatics Center, Professor in the Department of Clinical Research and Leadership at the George Washington University. She is an expert in research on text mining. She is also the Associate Director of the Center for Health and Aging at the Washington DC VA Medical Center. Dr. Zeng has over 20 years of experience in informatics, published over 80 peer-reviewed articles, and served as PI and Co-PI on a number of VA HSR&D, NIH, and DoD funded projects. Dr. Zeng serves as Co-Chair of the Women in AMIA CAC.
She will review the latest scientific research on work-life balance, including reported gender gaps (e.g. in compensation, tenure, publishing) in biomedical informatics and related fields.

Gretchen Purcell Jackson, MD, PhD, FACS, FACMI, FAMIA recently transitioned to a new position as Vice President and Chief Science Officer at IBM Watson Health and is an Associate Professor of Surgery, Pediatrics, and Biomedical Informatics at Vanderbilt University Medical Center since 2014. In her new role as an industry executive, she has faced new challenges in managing her family including more responsibilities and more travel for work. She has practiced as a pediatric surgeon, contributed as consumer health informatics researcher, a wife, a mother, and a stepmother. Dr. Jackson’s biomedical informatics research focuses on using technologies to empower patients and their families to take an active role in their health and healthcare. Dr. Jackson is a member of the AMIA Board of Directors and the Women in AMIA CAC. As a clinician and an executive, she will provide a unique perspective and discuss issues related to family planning, including career planning, work-work balance (clinical, research and education responsibilities), parental leave, caregiver guilt, and wellness.

William Hersh, MD, FACP, FACMI, FAMIA is Professor and Chair of the Department of Medical Informatics & Clinical Epidemiology in the School of Medicine at Oregon Health & Science University (OHSU). He is a leader and innovator in biomedical informatics in research and education. His research, in the area of information retrieval, includes over 200 scientific publications as well as the book, Information Retrieval: A Health and Biomedical Perspective, 3rd Edition (Springer, 2009). In education, he serves as Director of OHSU’s Biomedical Informatics Graduate Program, which is one of the oldest and largest in the world. Dr. Hersh is also involved in educational efforts in the physician clinical informatics subspecialty and also teaches clinical informatics to medical students, nursing students, graduate biomedical science students, and undergraduate health science students. There are four women faculty in his department who have career development (K) awards that he (and other faculty) mentor.

April Mohanty, MPH, PhD is Assistant Professor in the Department of Internal Medicine, at the University of Utah School of Medicine since 2018. She is also Health Research Scientist at the Salt Lake City VA, completed a Medical Informatics postdoctoral fellowship at the VA in 2017, and received a 5-year VA Career Development Award (a Historically Black Colleges and Universities-Research Scientist Training program) in 2016. Since 2017 she has Co-chaired the Women in AMIA CAC. As a Hispanic, 1st generation 4-yr college graduate, mother of two (toddler and newborn), and recently appointed faculty member, Dr. Mohanty will present her varied perspectives on some of the challenges and rewards related to becoming a new parent in academic and government settings as well as resources and advice for parents and minorities on navigating junior faculty member expectations.

Participation Statement I, April F. Mohanty, hereby confirm that all speakers have agreed to participate in this panel. Panelists are aware that there are no travel funds available. Panelists are also aware that the Working Groups are unable to reimburse registration costs.

References

6. Schulte B. Overwhelmed: How to work, love, and play when no one has the time. Macmillan, 2015.
Flipping Clinical Documentation on its Head


1Vanderbilt University Medical Center, Nashville, TN; 2OpenNotes/Beth Israel Deaconess Medical Center, Boston, MA; 3Stanford University School of Medicine, Palo Alto, CA; 4American Medical Informatics Association, Bethesda, MD

Abstract

Electronic health records (EHRs) have become ubiquitous tools and represent the standard of care for 96% of hospitals and 86% of ambulatory physicians in the United States. With adoption of EHRs came the promise of improved efficiency, higher quality care, and lower costs. Unfortunately, some clinicians are now spending twice as much time on documentation as they spend seeing patients, and the documentation paradigm of problem-oriented medical records is contributing to this imbalance. It is time to consider new innovations. In this panel, we present several ways that we can overturn the current paradigm, with a focus on increased readability, shareability, and information channels other than raw text. Clinical documentation serves many masters and exists within a complex regulatory environment, but we will tackle the implications of change head-on. We anticipate that this panel will be of interest to innovators and disruptors across the care continuum.

Learning Objectives: Upon completion of this panel, attendees will be able to:

- Understand why the current clinical documentation paradigm exists, and what can be done to improve it
- Enumerate some strengths and weakness of the collaborative wiki for clinical documentation
- Consider ways in which patients can interact with documentation, including co-editing
- Anticipate the role that voice-based agents will play in both the information retrieval and data creation stages of clinical care
- Name at least three examples of where visualizations can be preferable to written documentation

Panel Description

This interactive panel will present several new innovations with a common intent: changing the current paradigm of clinical documentation. Clinical documentation has been named by the National Academy of Medicine as one of the prime drivers of clinician burnout, which is estimated to affect up to 50% of those in the profession.1 The current method of encounter-based documentation is the evolution of several decades of policy, which have had the unintended effect of stymying innovation. In fact, some consider the Subjective/Objective/Assessment/Plan (“SOAP”) note introduced by Larry Weed M.D., in the late 1960’s to be the last true innovation in clinical documentation.2 In this panel we aim to turn common convention on its head through the introduction of several new innovations in this space. There will be one moderator and four speakers, as follows:

- **Jeffery Smith M.P.P.** (Vice President of Public Policy, AMIA) will moderate the panel. Jeff is responsible for planning, initiating, and managing AMIA’s overall public policy and government relations programs. Working with members, stakeholders, AMIA staff and relevant external organizations, Jeff consistently develops, communicates and disseminates AMIA’s public policy positions to advance the field of biomedical and health informatics. Specific to this panel, Jeff will discuss and field questions related to new and emerging policy directions pertinent to clinical documentation.

- **Jeremy L. Warner M.D., M.S.** (Associate Professor of Medicine and Biomedical Informatics, Vanderbilt University) will present “Collaboration, not Compendium: a Wiki-based Approach to Clinical Documentation”. In this presentation, he will introduce a new proposal for clinical documentation using the wiki format, based on a Perspective published in Academic Medicine.3
The collaborative wiki format offers many opportunities to ease the burden of documentation as well as to increase the usefulness of the recorded clinical data. Wikis support multiple authorship, have built-in features to track edits and changes, allow for contextual linkages (e.g., linking medical problems to their treatment), and support new technologies such as application programming interfaces, which allow for safe and secure exchange of information.

- **Liz Salmi** (Senior Strategist, Outreach & Communication, OpenNotes) will present “Our Health, Our Notes: Patients as Readers, Co-Authors, and Agenda Setters”. As a patient advocate and brain cancer survivor, Ms. Salmi has direct experience with the frustrations around access to notes. Once she obtained access, she became an avid reader of notes and quickly realized the benefits of reading, but also limitations in patient portal user experience. Now, through her work with OpenNotes, she is beginning to see a paradigm shift in both access and content of notes. A nascent initiative of OpenNotes, which is an international movement advocating fundamental change in the way visit notes are managed, is to put patient’s in the driver’s seat. This includes new innovations in shared note authorship, as well as pre-visit agenda setting by patients.

- **Yaa Kumah-Crystal M.D., M.P.H., M.A.** (Assistant Professor of Biomedical Informatics and Pediatric Endocrinology, Vanderbilt University) will present “Let’s not Mince Words: Voice as the Means to Read and Write Notes”. Apple has Siri, Google has Google Assistant, and Amazon has Alexa. A team at Vanderbilt University Medical Center (VUMC) led by Dr. Kumah-Crystal is creating voice-controlled virtual assistant to help people interact with electronic health records (EHR) using natural speech. They call their project VEVA, the Vanderbilt EHR Voice Assistant. VEVA represents a new way to interface with the medical record to collect and concatenate data, and in the future will also work bidirectionally as a content generator. The future of natural language processing in medicine could mean virtual assistants functioning as virtual scribes. Voice is a natural medium of communication that may help re-humanize the EHR.

- **Chethan Sarabu M.D.** (Clinical Instructor of Pediatrics, Stanford Medicine) will present “Show, Don’t Tell: How Visuals Improve Healthcare Visits”. Dr. Sarabu recently completed his fellowship in clinical informatics and his primary focus was around laying the groundwork for collaborative documentation in different ways from OpenNotes in pediatric populations to patient questionnaires populating note templates. His presentation will challenge the predominance of text-based documentation of verbal conversation. There is a need to bring more visuals into documentation, such as timelines, photographs, or even emoji to communicate information in ways that can provide rich meaning with brevity. Furthermore visuals can bring down communication barriers in populations such as children and non-native English speakers. The use of visual narrative in documentation may be especially powerful for patients with complex longitudinal diseases, as well as for those who learn best from methods other than text-based.

**Intended Audience:** We anticipate that this panel will be of interest to a wide range of individuals attending the AMIA Annual Symposium. Specifically the following: (1) Clinicians who create clinical documentation; (2) Patients who have an interests in participatory medicine including co-authorship of clinical documents; (3) Clinical researchers and informaticians interested in extracting information from healthcare data; (4) Policymakers who are considering changes to the current paradigm of clinical documentation; (5) Students and trainees interested in learning more about the past, present, and future of clinical documentation.

**Timeliness**

We feel that this panel could not come at a better time. After decades of static policies, federal agencies have recently proposed changes to the way clinicians must document care. Specifically, the Centers for Medicare & Medicaid Services (CMS) has committed to E/M documentation guideline reform, which will take effect in 2021, allowing practitioners greater flexibility to exercise clinical judgment in documentation. Further, CMS and the Health & Human Services’ Office of the National Coordinator for
Health Information Technology (IT) have published a sprawling strategy to reduce health IT-related regulatory and administrative burden. When considered alongside broader changes to the health IT landscape, these federal efforts provide stakeholders with a unique opportunity to consider alternatives to long-standing documentation dogma.

**Statement from the Panel Organizers**

Dr. Warner, the panel organizer, affirms that all panel participants have agreed to participate and have contributed to the preparation of this document (as of March 9, 2019).

**References**

3. Warner JL, Smith J, Wright A. It’s time to wikify clinical documentation: how collaborative authorship can reduce the burden and improve the quality of the electronic health record. *Acad Med* 2019;Published Ahead of Print: https://doi.org/10.1097/ACM.0000000000002613.
The Nuts and Bolts of Sponsorship: Perspectives in Informatics

Tiffany I. Leung, MD, MPH, FACP, FAMIA¹, Marion J. Ball, EdD, FAAN, FACMI²,³, Cynthia Brandt, MD, MPH⁴,⁵, Gretchen Purcell Jackson, MD, PhD, FACS, FACMI, FAMIA⁶,⁷, Titus Schleyer, DMD, PhD, FACMI, FAMIA⁸,⁹

¹Maastricht University, Maastricht, The Netherlands; ²Center for Computational Health, IBM Research; ³Johns Hopkins University School of Nursing, Johns Hopkins School of Medicine, Baltimore, MD; ⁴VA Connecticut Healthcare System, West Haven, CT; ⁵Yale University School of Medicine, New Haven, CT; ⁶IBM Watson Health, Cambridge, MA; ⁷Vanderbilt University Medical Center, Nashville, TN; ⁸Indiana University School of Medicine, Indianapolis, IN; ⁹Center for Biomedical Informatics, Regenstrief Institute, Inc, Indianapolis, IN

Abstract

Sponsorship is important for career development, but infrequently discussed and often not well understood. Sponsors are advocates for their protégés, recommending them for promotion, connecting them with senior leaders, promoting their visibility – pulling them up the ladder with them¹. In this panel, the speakers will introduce the concept of sponsorship, its impact on careers of young, mid- and experienced informaticians, and describe their own experiences from both sides of the sponsorship relationship. The objectives of this panel are to: (1) learn the importance of sponsorship in one’s career; (2) learn how to be a sponsor or to find one; (3) share experiences with others; and (4) develop an action plan for what to do during the AMIA symposium to put into practice what is learned.

Introduction

Sponsorship is as important for career development as mentorship, but less often discussed and often not as well understood. Being a sponsor may mean being well-connected and well-established in a field and a talent scout, while being a protégé may mean being proactive, loyal, and focused on specific opportunities ², such as promotion, award nomination or recognition, and other specific episodes in career progression. Sponsors are advocates for their protégés, recommending them for promotion, connecting them with senior leaders, promoting their visibility – pulling them up the ladder with them¹,³. In essence, sponsorship aims to enhance the visibility and credibility of talented individuals⁴. There is also growing attention to the importance of diversity and inclusivity in science; sponsorship can be a pathway towards overcoming biases that disfavor individuals based on ethnicity, race, gender orientation, health or disability status, age, beliefs, religion, and other characteristics⁵,⁶. However, this also means deliberate practice of sponsorship, as a sponsor or a protégé, must be intentional and focused in order to benefit individuals and their organizations.

In this interactive panel, sponsored by the Women in AMIA Subcommittee on Awards & Leadership, the speakers will introduce the concept of sponsorship, its impact on careers of young, mid- and experienced informaticians, and describe their own experiences and lessons learned from both sides of the sponsorship relationship. The objectives of this panel are to: (1) learn the importance of sponsorship in one’s career; (2) learn how to be a sponsor or to find one; (3) share experiences with others (through small group and pair-and-share activities); and (4) develop an action plan for what to do during the AMIA symposium to put into practice what is learned (individual reflection). Audience participants’ comments and panelists’ presentation content will be synthesized and shared electronically after the session to provide a reference for sponsorship best practices.

<table>
<thead>
<tr>
<th>Session Agenda</th>
<th>Description</th>
<th>Presenter</th>
<th>Duration</th>
</tr>
</thead>
</table>
| Introductions & What is Sponsorship? | • Objectives and Structure of Panel  
• Panelist introductions  
• What is sponsorship? Audience poll for live word cloud.  
• Objectives & Definitions | Leung (moderator) | 10 minutes |
| Sponsorship vs. Mentorship | How are sponsorship and mentorship similar or different?  
How does sponsorship specifically help career progression in ways that mentorship does not? | Ball | 10 minutes |
### Research & Recognition

Sponsoring or receiving sponsorship for *research and recognition.*

**Schleyer**
10 minutes

#### Interactive Component: Pair & Share or Small Groups of 4 or less

Discuss in pairs or groups of 4 or fewer individuals one of these two perspectives:*  
- As a sponsor, *how would you identify those to sponsor?*  
- As a sponsor, *how would you know that you could be a sponsor?*  
- As a protégé (recipient of sponsorship), *how can one be proactive about connecting with a sponsor? For what purpose(s) would you look for sponsorship?*  
- As a protégé, *how would you ask for sponsorship?*

**All speakers**
10-15 minutes

#### Interactive Component: Report out

Pairs or small groups share their key discussion points, which will be compiled into a shareable document.*

**All speakers**
10 minutes

#### Job Transitions & Changes

How does sponsorship support navigating the “jungle gym” of career advancement within and between academics, industry, etc?

**Jackson**
10 minutes

#### Diversity & inclusivity

How is sponsorship as aligned with principles of diversity and inclusivity?

**Brandt**
10 minutes

#### Interactive Component: Self-Reflection

Each participant reflects on the content of the panel and write down one thing they will do during the AMIA Symposium or within the next 1 week to *seek or be a sponsor* for one of their professional goals. Goal should be specific:  
- *Who will you engage as a sponsor/protégé?*  
- *What is the aim of such sponsorship (e.g. recognition, promotion, etc)?*  
- *What action(s) will you take towards that goal? When?*

**Leung**
5 minutes

#### Closing

Moderator summary  
Open discussion, during which participants may share their self-reflection tasks or make other comments, and Q&A

**Leung**
10 minutes

* Panelists and members of the Women in AMIA Subcommittee on Leadership and Awards who are in the audience may engage the small groups also in discussion, facilitate their sharing with the larger audience.

### Possible Discussion Questions

Questions to be used as prompts for the interactive components of this panel are detailed as above.

### Panelist Biographies

**Tiffany I. Leung, MD, MPH, FACP, FAMIA** (moderator) is a U.S.-trained Internal Medicine physician, Assistant Professor and also a PhD candidate at Maastricht University in the Netherlands. Her current work focuses on implementing value-based health care at Maastricht University Medical Center +; her research examines value-based health care redesign using a systems thinking approach. Dr. Leung is also studying healthcare professionals’ competency and professionalism development in the context of system redesign, and aims to design interventions to overcome knowledge and skill boundaries among workers engaged in such work. She recently took on the role of chair of the Women in AMIA Subcommittee on Awards & Leadership, succeeding Jessie Tenenbaum, PhD, FACMI.

**Marion J. Ball, EdD, FCHIME, FHIMSS, FMLA, FAHIMA, IAHSI** is a member of the National Academy of Medicine (NAM). She currently serves on the Board of Health on the Net (HON) in Geneva, Switzerland, was elected to membership of the IBM Industry Academy. She has also served on a variety of boards in the area of health information technology: the Board of Regents of the National Library of Medicine (NLM), as President of the International Medical Informatics Association (IMIA), as a board member AMIA, CHIME, AHIMA, and Co-Chaired...
the HIMSS Board, and also served as Board Co-Chair of the TIGER Foundation. She received the Morris F. Collen Lifetime Achievement Award from ACM/AMIA, and is an honorary member of Sigma Theta Tau, the Honor Society of Nursing, and the Medical Library Association (MLA). She received the Award of Excellence, an International Medical Informatics Association (IMIA) Lifetime Achievement award. In 2017, she was recognized by HIMSS as one of The Most Influential Women in Health IT, and also named as one of the Most Powerful Women in Healthcare IT by Health Data Management. She is the author/editor of over 25 books and over 230 articles in the field of Health Informatics. Her book, Consumer Informatics, received the HIMSS Book-of-the-Year Award. Several previous editions of the nursing books have been translated into other languages.

**Cynthia Brandt, MD, MPH** is the Director for the Yale Center for Medical Informatics, Professor of Emergency Medicine, Anesthesiology and Biostatistics at Yale Schools of Medicine and Public Health, and Director of the Informatics Core for the Center of Innovation at VA Connecticut. Dr. Brandt completed a general Preventive Medicine residency at Madigan Army Medical Center in 1989 and a postdoctoral fellowship sponsored by the National Library of Medicine at Yale School of Medicine in 1997. She is board certified in Preventive Medicine and Clinical Informatics. Her research is interdisciplinary and focuses on issues related to the design, development and use of informatics tools in the domain of clinical research, as well as health services research. She is PI of the National Library of Medicine T15 Training Grant at Yale, and Co-Director of the Medical Informatics and Women’s Health Research Fellowships at VA Connecticut.

**Gretchen Purcell Jackson, MD, PhD, FACS, FACMI, FAMIA** is Vice President and Chief Science Officer at IBM Watson Health and an Associate Professor of Surgery, Pediatrics, and Biomedical Informatics at the Vanderbilt University Medical Center. She is an internationally recognized informatician and surgeon with accomplishments spanning over 25 years of. Dr. Jackson is a member of the AMIA Board of Directors. She earned her PhD in Medical Informatics at Stanford University in 1997 and held informatics faculty and leadership positions at three universities during surgical residency and fellowship. Dr. Jackson is a member of the editorial boards of the *Journal of Biomedical Informatics* and *International Journal of Medical Informatics*, and an associate editor of *Applied Clinical Informatics* and *JAMIA Open*. Her current research focuses on evaluating clinical decision support systems and empowering patients and families through health information technologies.

**Titus Schleyer, DMD, PhD, FACMI, FAMIA** is Professor of Biomedical Informatics in the Department of Medicine, Indiana University School of Medicine, as well as a Research Scientist at the Center for Biomedical Informatics at the Regenstrief Institute. Primarily funded by the National Institutes of Health, Dr. Schleyer's current research is focused on making information from various sources more accessible in clinicians' workflows, big data mining for population health management and research, developing analytics and visualizations to address the opioid epidemic, and the development of the Indiana Learning Health Systems Initiative. He co-directs (with Dr. Brian Dixon) the National Library of Medicine-funded Public & Population Health Informatics (PHI) Fellowship at Indiana University.

**Participation Statement**

All participants have agreed to take part on the panel.

**References**

3. Harris C. How to find the person who can help you get ahead at work. [cited 2019 Mar 6]. Available from: https://www.ted.com/talks/carla_harris_how_to_find_the_person_who_can_help_you_get_ahead_at_work?language=en
Building the Health Apps Economy under 21st Century Cures: Interplay among EHRs, Third Party Apps, and the Cloud

Panel Moderator

Kenneth D. Mandl, MD, MPH
Computational Health Informatics Program, Boston Children’s Hospital
Department of Biomedical Informatics, Harvard Medical School
Boston, MA USA

Panelists

Aashima Gupta
Google Cloud
Mountain View, CA USA

Josh Mandel, MD
Microsoft Healthcare
Redmond, WA, USA
Department of Biomedical Informatics, Harvard Medical School
Computational Health Informatics Program, Boston Children’s Hospital
Boston, MA USA

Matt Obenhaus
Cerner Corporation
Kansas City, MO USA

Don Rucker, MD
Office of the National Coordinator of Health Information Technology
Department of Health and Human Services
Washington, DC USA

Isaac Vetter
Epic
Madison, WI USA
**Introduction of the Topic.** A decade after the Meaningful Use Program was first established, over 95% of inpatient settings use certified electronic health records (EHRs) and over 80% of office-based practices use certified EHRs. We are now entering the next phase of health IT development, as a clear pathway toward interoperability among different systems emerges—a universal, standards-based, public application programming interface (API). The 21st Century Cures Act, a bill that enjoyed strong bipartisan support which was signed into law at the end of 2016, requires that certified health IT products have an API that allows all of a patient’s health information to be accessed, exchanged, and used “without special effort.” This requirement follows from years of advances and real world implementations under the SMART on FHIR project and development and maturation of the Fast Health Interoperability Resources (FHIR) standard.

The panel is extremely timely—two new HHS draft rules were recently released, following on from the Cures Act, firmly establish FHIR as the lingua franca for health data representation and SMART as the API between health data in certified systems and the innovators that use those data. At the time of this panel, the public comment period will have ended and final rule-making will be in process.

Under the Argonaut consensus working group, all the major EHR vendors have implemented the SMART API in their products. To date, the two most impactful SMART on FHIR implementations are the Apple Health App, which connects to hundreds of health systems via the SMART API, and the CMS Blue Button 2.0, which uses the SMART API to provision claims data to its beneficiaries. The total number of SMART apps in the various public and commercial app galleries, however, have been relatively modest.

Possibilities for the resulting ecosystem, and the dynamics underpinning it, are just beginning now to come into focus. It is widely believed that genomic medicine and artificial intelligence will enter patient and physician workflows through SMART on FHIR apps. With the regulated uniformity in how clinical data are represented and how apps connect to health IT, we expect that billions in capital will be invested in extending the digital, AI, and genomics capabilities of healthcare well beyond where we are today.

**Intended Audience.** The panel is directed broadly at the AMIA membership, especially those focused on biomedical informatics from a clinical translational perspective. We expect it will be of particular interest to those engaged in genomic medicine, machine learning, decision support, app development, enterprise IT, and EHR innovation.

**Aim of the Discussion.** This panel brings together leaders representing public API development, the government agency spearheading health IT interoperability, the two largest EHR vendors, as well as relatively new entrants into the space—major cloud services companies. This interactive discussion will engage the audience in exploration of barriers to and opportunities for creating a burgeoning health apps economy, one that accelerates the translation of innovation to care, well established in provider workflows.

**Specific Contribution of Each Speaker.**

The moderator, and each speaker will offer framing remarks for 5 minutes and then engage with the audience in a brisk, interactive, moderated discussion. The moderator will keep the flow moving fast, alternately posting questions to the panelists and taking questions from the audience.

**Ken Mandl,** Director of the Computational Health Informatics Program at Boston Children’s Hospital and a founder of the SMART Health IT project, will moderate the panel. He will begin with a brief overview of the current state and present a scenario of a hospital attempting to both purchase third party apps, but also to internally develop and deploy SMART apps. The use case will outline opportunities to use native EHR FHIR APIs as well as “side car” approaches. Challenges to physician workflow, app validation, and commercial success of third party apps will be highlighted.
Don Rucker, National Coordinator of Health Information Technology, will discuss the Administration’s strategy in promoting interoperability through the Cures API and information blocking provisions.

Aashima Gupta, Director of Global Healthcare Solutions for Google Cloud, will discuss their FHIR cloud strategy, including its offering of the SMART API. She will address approaches and challenges to offering secure, turnkey solutions for API-enabled FHIR servers.

Isaac Vetter, Integration Lead for the Epic App Orchard will present current and upcoming technologies which enable healthcare apps, considerations for protecting security & privacy and the evolution of Epic's App Orchard as shaped by health system & developer feedback as well as regulation.

Josh Mandel, Lead Architect for SMART Health IT and Chief Architect at Microsoft Healthcare will discuss the challenges of scaling an app ecosystem across Health IT, especially from the perspective of open specifications and standards. He will describe challenges of consensus, outlining the role of open source projects; small communities focused on specific problems; and the wider, formal standards development process. He will describe the pivotal role of open standards in Microsoft's healthcare cloud computing strategy.

Matt Obenhaus, Director and Solution Executive for Open Developer Experience at Cerner, will discuss the Cerner App Gallery and evolving models to encourage developers to create value. He will also discuss a roadmap for exposing new data types across the SMART API on Cerner.

Expected Discussion. At a pivotal time for change and progress, the audience will be brought to the cutting edge of progress and decision making in the largest health IT companies in the world, and in the federal agency that both promotes health IT and regulates it. Who will produce the next generation of EHR-connected apps. Where will the now regulated APIs sit? How will workflows develop across EHRs and Cloud hosted side cars? Audience members will no doubt be key informants to the discussion as many will have used the open APIs, deployed apps, and begun to creatively use the cloud services.

Sample topics for discussion:
- Will EHR vendors rely on “side car” FHIR servers with near real time data for app functionality as well?
- What are plans for triggering decision support, e.g., CDS Hooks?
- What is the vision for app reusability—an app written once will run anywhere in healthcare?
- What health IT services do the cloud vendors envision offering?

Panel Organizer Statement. All participants have agreed to take part on the panel.
Results from a Multi-stakeholder Action Plan to Better Leverage Patient-centered Clinical Decision Support in Addressing the Opioid Misuse Crisis

Jerome A. Osheroff, MD\textsuperscript{1}, Craig Robbins\textsuperscript{2}, MD, MPH, Brian Alper, MD, MSPH\textsuperscript{3}, David Little, MD, MS\textsuperscript{4}, Edwin A. Lomotan, MD\textsuperscript{5}
\textsuperscript{1}TMIT Consulting, LLC, Naples, FL; \textsuperscript{2}Kaiser Permanente, Denver, CO; \textsuperscript{3}EBSCO Health, Ipswich, MA; \textsuperscript{4}Epic, Verona, WI; \textsuperscript{5}Agency for Healthcare Research and Quality, Rockville, MD

Abstract

Over 50 individuals from many stakeholder groups convened in 2018 to develop and begin executing an action plan to improve pain management, opioid use, and opioid use disorder treatment through more widespread and effective patient-centered clinical decision support (PCCDS). The opioid action plan (OAP) includes five consensus, highly desirable opioid-related future vision scenarios for PCCDS-enabled care. The scenarios reference 19 PCCDS interventions that are not widely or well implemented today but could be highly valuable. The OAP includes recommendations for various stakeholder groups to make the future vision and underlying interventions widely implemented. Most importantly, the OAP includes over two dozen examples of specific steps toward the future vision that organizations have taken as a result of contributing to the OAP. During this panel, the OAP lead author will outline the future scenarios and recommendations, and 4 individuals who helped develop the plan will describe OAP-driven actions they have taken and/or implications for their work with healthcare systems, CDS suppliers, electronic health record vendors, and government agencies. Half the session will be interactive discussion between panelists and audience members to surface opportunities to further accelerate PCCDS-enabled pain and opioid management within and across organizations.

Introduction

The United States is experiencing an opioid misuse and overdose epidemic that results in staggering costs, morbidity, and mortality and has caused a national crisis\textsuperscript{1}. Poorly managing pain - a highly prevalent and distressing symptom - is a key underlying driver for this epidemic\textsuperscript{2}, as is the underuse of effective treatments for opioid use disorder (OUD)\textsuperscript{3}.

Evidence-based guidance is available for pain management, opioid use and OUD treatment\textsuperscript{4}, though approaches to putting this information into practice are suboptimal. Clinical decision support (CDS) can enhance health-related decisions, actions and outcomes by delivering the right information (e.g., evidence-based guidance and tools) to the right people, in the right formats through the right channels at the right times\textsuperscript{5}. CDS strategies enabled by modern healthcare information technology offer many targeted opportunities to provide information when, where and how it’s needed to optimize patient and care team decisions, actions and partnership.

“Patient-centered Clinical Decision Support” (PCCDS) is a CDS subtype that focuses on highly patient-focused strategies, interventions and tools for enhancing communications, decisions and actions by and between patients and clinicians\textsuperscript{6}. For example, PCCDS interventions for pain and opioid-related care include tools such as smartphone apps that patients can use to help them track and manage pain and function according to a care plan developed with their care team, shared decision-making tools for selecting pain management strategies, and screening tools to detect patients at high OUD risk.

A PCCDS Opioid Action Plan (OAP) was produced as a central 2018 project of the Patient-Centered Clinical Decision Support Learning Network (Learning Network), an Agency for Healthcare Research and Quality (AHRQ)-funded initiative to more fully leverage PCCDS in making care more patient centered while broadly improving care delivery and outcomes\textsuperscript{7}.

Developing and Executing the PCCDS OAP

The Learning Network chartered an OAP Workgroup (OAPWG) in April 2018 to develop the OAP. This group grew to include over 50 individuals from many stakeholder groups and met by teleconference 17 times between April and December 2018. They also worked on OAP elements using online collaboration tools (Google docs) and the OAP was a focus of the Learning Network’s Annual Meeting in October 2018. Figure 1 below illustrates the various stakeholder groups and individual organizations represented in the OAPWG or that otherwise supported OAP development.
Figure 1: Stakeholder Groups and Interactions (Shaded boxes represent groups most directly involved in realizing the future vision scenarios; CDO=care delivery organization, PHM=population health management systems, CDSS=clinical decision support systems; EHRs=Electronic Health Records)

Key OAP elements the OAPWG developed included a consensus future vision of PCCDS-enabled pain management, opioid use, and OUD treatment; an initial survey of current resources and activities that could be leveraged to achieve the future vision; recommendations for different stakeholder groups to drive progress toward this vision; and actions triggered by the OAP development process that begin implementing these recommendations.

Figure 2 below outlines the topic areas for the five consensus future vision scenarios that the OAPWG developed.

Figure 2: Overview of Future Vision Scenario Focus Areas

The scenarios are each about a page long and outline PCCDS-enabled workflow and information flow that follows patients across settings – e.g., self-care at home, office visit, and hospitalization. The scenarios contain personas with different patient ages, genders, and clinical conditions. Even though they are specific in this way, the scenarios exemplify very broad opportunities for PCCDS to better support pain management, opioid use and OUD treatment.

As part of outlining desirable workflows and information flow for each clinical scenario, the OAPWG defined specific PCCDS interventions that could be used to implement these enhanced care processes. These interventions include care plan development and tracking tools, condition-specific pain/opioid-related order sets, OUD detection and notification algorithms, dashboards outlining patient-specific pain management treatments and results, shared decision-making tools, and many others. The scenarios outline where and how these interventions support care.

Another key OAP element is recommendations for various stakeholder groups (i.e., as outlined in Figure 1) to make the future scenarios – enabled by the underlying PCCDS interventions - widely implemented and highly valuable. These recommendations include steps to ensure that health IT suppliers provide the needed interventions, that providers implement these tools successfully in workflow to realize the future scenarios, that patients likewise use the interventions independently and with their care teams to achieve the desired processes and outcomes, and that all the other stakeholder groups support these core activities as appropriate.
Most importantly, the OAP was developed in an action focused way, and it includes over two dozen examples of specific steps that organizations have taken as a result of their engagement in developing the OAP to drive progress toward implementing the recommendations and achieving the desired future vision scenarios.

**Proposed Panel**

The goal of this panel is to build on the valuable dialog, activities, and collaboration reflected in the OAP to benefit more organizations from this work and accelerate widespread progress toward the future vision. Annual Symposium attendees typically reflect many of the stakeholder groups listed in Figure 1, and this panel is aimed at such participants that have pain management, opioid use and/or OUD treatment as a focus.

After a brief OAP overview, JAO will describe the future vision scenarios and underlying PCCDS interventions to stimulate attendees to reflect on whether/how their professional and organizational goals can be supported by leveraging them as guideposts for their efforts. He’ll also briefly summarize the recommendations and OAP-driven actions to implement the scenarios as a model for attendee action. CR will describe actions that Kaiser Permanente has taken to execute the OAP, e.g., sharing tools they’ve developed and exploring collaborations with other panelists to develop new, integrated PCCDS tools to realize the future scenarios. BSA will describe new offerings that EBSCO created as a result of his OAPWG participation – i.e., a new patient-facing tool with evidence-based guidance for selecting between opioid and nonopioid alternatives for common chronic pain scenarios. BSA will share implications more broadly for CDS suppliers and international consortiums working on standards for interoperability. DL will describe how Epic has used the future scenarios to inform its product development and implementation support team efforts to provide more powerful CDS opioid-related CDS and ensure clients fully leverage the pertinent capabilities already available. He will also share activities and implications more broadly for EHR vendors and the EHR Association (EHRA). EAL will describe how the OAP fits in AHRQ’s CDS portfolio and implications for OAP execution.

The panel’s second half will be devoted to interactive discussion among panelists and audience members to surface opportunities to further accelerate PCCDS-enabled pain management and opioid use within and across individual organizations. Prompts for this discussion include: Do the future scenarios seem valuable for your efforts? If so, how? Are there ways your organization can benefit from actions/collaborations the panelists pursued? If so, what/how? How could any subsequent activities to execute the OAP most benefit your organization? Are there ways you or your organization could contribute to further OAP execution?

All participants have agreed to take part in the panel.

**Conclusion**

Patient-centered clinical decision support is a promising but underutilized approach to addressing the national opioid misuse and overdose crisis. A broad, multi-stakeholder initiative has outlined a shared future vision for broader and more effective PCCDS use for this target, and many organizations have already started implementing recommendations for achieving this vision as part of developing the OAP. This interactive panel will build on this work and leave attendees with valuable insights and opportunities to advance their efforts in this area.

**References**


Actionable opportunities for improving knowledge at intersection of dental and medical data

Neel Shimpi BDS, MM, PhD; Karmen S. Williams DrPH, MBA; Alexander V. Alekseyenko, PhD

Moderator: Alexander Alekseyenko
Speakers: Chad Novince DDS, MSD, PhD
          Amit Acharya BDS, MM, PhD
          Muhammad Walji MS, PhD
          Thankam Thyvalikakath DMD, MDS, PhD

1 Center for Oral and Systemic Health, Marshfield Clinic Research Institute, Marshfield, WI, USA
2 Indiana University School of Public Health, Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, USA
3 Program for Human Microbiome Research, Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA
4 Department of Oral Health Sciences, Medical University of South Carolina, Charleston, SC, USA
5 Diagnostics and Biomedical Sciences, Technology Services and Informatics, UTHealth, University of Texas, Houston, TX, USA
6 Cardiology, Operative Dentistry & Dental Public Health, Indiana University School of Dentistry, Indianapolis, IN USA

Abstract:

Advances in the understanding of oral and systemic connections and a paradigm shift in delivering person-centered care has created a demand for incorporating principles of interprofessional education and healthcare delivery. Historically, dentistry and medicine have evolved as separate domains, and this has influenced education and healthcare delivery. Lack of interprofessional training during professional school critically influences practice patterns of dental and medical providers. Interprofessional health information exchange and integrated medical-dental electronic health records are critically vital to integrated care delivery. The lack has raised challenges as well as opportunities for establishing informatics infrastructure to support interdisciplinary care delivery in the context of oral-systemic health conditions. The purpose of this panel is to provide considerations for the establishment of informatics infrastructure and required training of the integrated care delivery team. Our goal is to achieve a roadmap for alignment of informatics and interprofessional efforts to maximize the broad positive impact on continuity of care.

Brief description of each panelist’s presentation:

- **Dr. Chad Novince**, “Need for the development/integration of interprofessional curriculum across medical schools and dental schools”
  This presentation will focus on how interprofessional healthcare knowledge and patient care experiences during professional school critically influence whether dentists and physicians will practice interprofessional healthcare post training. Interprofessional medical – dental curriculum could be designed in various formats, but importantly should include both didactic and clinical interprofessional training experiences.

- **Dr. Amit Acharya**, “Beyond the Integrated Medical-Dental Electronic Health Records”
  This presentation will discuss how a Healthcare System in Wisconsin has applied an integrated medical-dental EHR platform to solving many medical-dental practice divide. The topics will include a, point of care clinical decision support to better manage conditions with oral-systemic associations, b, secondary use of clinical data for
research: how routinely collected clinical data from the integrated medical-dental EHR platform is enabling research to better understand the several oral-systemic associations, and c. development of quality improvement measures to better support the management of conditions with oral and systemic associations.

- **Dr. Muhammad Walji. “Towards a Person-centered care Learning Health System”**
  This presentation will discuss the concept of building a sustainable dental learning health system focused on continuous quality improvement and providing patients with safe and effective oral healthcare. The presentation will highlight how a group of large academic health systems are working together to setup a practice-level learning health system supported by an effective informatics infrastructure, interprofessional collaborations, relevant data governance and a culture allowing for self-reflection and learning.

- **Dr. Thankam Thyvalikakath. “Leveraging patient information through regional health information exchange to improve oral health”**
  This presentation will discuss the process of matching dental patient data with their medical record data available through a regional health information exchange to promote continuity of care across medical and dental care settings. Dr. Thyvalikakath will describe through a case study that determined the Sjögren’s Syndrome diagnosis of dental patients and the influence of dental clinicians’ awareness of their patients’ diagnosis on managing oral health.

**Relevance of topic and anticipated audience**

The oral and systemic health has been a priority area for the NIH National Institute for Dental and Craniofacial research. Informaticians and data scientists will need to learn what opportunities exist to improve health by integration of dental and health medical record data. Healthcare providers including dental and medical providers will need to expand their understanding of the impact of their practice on other areas of their patients’ health and how to improve their own treatment decisions by holistic understanding of their patient. Educators will learn how to incorporate interprofessional education paradigms to cross-train physicians and dentists in essential concepts of both areas. Finally, the researchers will learn about the resources that can drive further understanding between oral and systemic health.

**Discussion questions:**

1. How does interprofessional education influence practitioners’ delivery of care?
2. What translational value does an integrated care system hold?
3. What are the technical and non-technical barriers for implementing a learning health system?
4. Why is it important to practice in an interdisciplinary environment?

**Learning Objectives:**

1. Identify how to capitalize on the opportunities for developing an integrated medical-dental curriculum that will support interprofessional healthcare delivery.
2. Understand and recognize the potential application of an integrated medical-dental EHR platform in solving several medical-dental delivery system challenges and further identifying translational areas for establishing integrated models of care.
3. Understand the key components of developing a multi-institutional learning health system.
4. Learn the potential of utilizing regional health information exchange to enhance continuity of care across dental and medical care settings.

The organizer members confirm that all participants have agreed to take part on the panel.
Creating a Learning Health System for Measurement of Diagnostic Safety: Emerging Implications for Health Information Technology

Hardeep Singh, MD, MPH1, Dean Sittig, PhD2, Ashley N.D. Meyer, PhD1, Traber D. Giardina, PhD, MSW1 and Divvy Upadhyay, MBBS, MPH3

1Michael E. DeBakey Veterans Affairs Medical Center and Baylor College of Medicine, Houston, TX; 2University of Texas Health Science Center, Houston, TX; 3Geisinger, Danville, PA

Abstract

Diagnostic errors pose a significant threat to patient safety.1 Learning health systems (LHS) provide opportunities to investigate, monitor, and measure diagnostic errors in real time within a healthcare system. A multidisciplinary panel will discuss emerging priorities and research opportunities related to health information technology (HIT)-enabled measurement of diagnostic errors, now considered the next big patient safety challenge.2 This session will explore how to use multiple sources of data from the health system (electronic triggers), providers, and patients, to identify and analyze missed opportunities in diagnosis. These data can be used to understand how health systems can learn and enhance the safety and accuracy of the diagnostic process. The learning objectives of this panel include: 1) how HIT-enabled health care organizations can become LHS for improving diagnosis; and 2) how informaticists can help implement pragmatic approaches to measure and understand diagnostic errors and engage in development of institutional infrastructure and feedback processes to promote individual and organizational learning.

Panel Description

Diagnostic errors affect at least 5% of US adults every year,3 leading to an estimated 40,000-80,000 deaths yearly. Yet, learning health systems approaches for investigating causes and potential solutions to diagnostic error are underdeveloped. Measurement is the first step to understanding the problem scope. The National Academy of Medicine’s 2015 report, Improving Diagnosis in Health Care,1 recommended measurement strategies that use HIT and involve both patients and health care professionals. In the last decade, research in both learning health systems and informatics have advanced the science of measurement and the time to engage AMIA audience is ripe.

This panel will take a deep-dive into HIT-based measurement methods and concepts, and current gaps in measurement related to diagnostic quality and safety. This will provide a multidisciplinary perspective into what challenges and opportunities others might encounter as they attempt to delve into the messy world of diagnostic error research, implementation and improvement activities. Because diagnostic quality and safety is still an emerging area of measurement, we will discuss current informatics research and potential measurement strategies in the three broad domains: provider/care team perspectives on use of HIT tools to find and reduce diagnostic errors; health care delivery system and policy perspectives aimed at using HIT to measure, monitor, and intervene to prevent diagnostic errors; and patient/family/caregiver perspectives on use of HIT to improve communication with providers and health systems about diagnosis-related care gaps. We will include discussion of studies related to activating the patient’s role in monitoring the diagnostic process. Strengths and weaknesses of various HIT-enabled measurement methods used in diagnostic safety research will be discussed, along with the importance of multidisciplinary research, the triangulation of research findings and methods, and partnering with clinicians, patients and health systems. We will also introduce more audience and panel interaction by using an audience response system called Slido, which would generate not only questions but also comments to address in Q&A. This content, as well as the interactive methods for delivering it, will contribute to more discussion, debate and audience engagement.

Panel Members

Hardeep Singh, MD, MPH (moderator, organizer, and panelist – 10 minutes) will start off the panel by introducing the panelists and the topic including its greater significance and context in the current healthcare landscape. Dr. Singh will also describe the rationale and motivation for bringing this diverse panel together.

Then, Dr. Singh will provide an overview of systems-based methods to measure diagnostic errors. He will discuss use of trigger tools to mine patient data to identify signals indicative of potential errors or adverse events. He will discuss the use of validated electronic trigger algorithms that analyze electronic health record (EHR) data to identify patients with potential diagnostic safety concerns. A newly developed data mining and knowledge discovery framework, the
Safer Dx Trigger Tools Framework, will be used to illustrate the application of electronic algorithms to identify targets for improving diagnosis.

Ashley N.D. Meyer, PhD (panelist– 10 minutes) will discuss provider-based measurement methods. She will discuss using diagnostic performance measures and other data to create and deliver actionable feedback to clinicians, allowing them to learn from and improve upon their diagnostic performance. Potential reluctance of clinicians to receive diagnostic performance feedback and other potential issues will be discussed, along with potential recommendations for providing feedback in a way that improves diagnosis and harmonizes with a learning health system.

Traber D. Giardina, PhD, MSW (panelist– 10 minutes) will discuss patient-based measurement methods. She will describe the value of learning from patient/family/caregiver reports of adverse events and using patient experience data to supplement current patient safety initiatives. She will discuss emerging challenges in this area and describe recently used methods for examining diagnostic errors from the patient perspective, including incident reporting and patient complaints. Dr. Giardina will discuss recent evidence from a patient-focused study and highlight possible measurement strategies that include the use of HIT to effectively capture patient experiences.

Divvy K. Upadhyay, MBBS, MPH (panelist– 10 minutes) will discuss challenges and barriers to LHS approaches in HIT-related measurement. Discussion will include early lessons from the Safer Dx Learning Lab, a unique partnership that uses multiple sources of data from the health system (risk management, electronic triggers), providers and patients to identify and analyze missed opportunities in diagnosis. The lab aims to understand how health systems can enhance the safety and accuracy of the diagnostic process and involves a close partnership between multidisciplinary diagnostic safety researchers, a pioneering health system, and a Safer Dx Researcher in Residence based at the health system.

Dean Sittig, PhD (discussant– 5 minutes) will put all the presentations in context of health care systems that now have access to increasing amounts of longitudinal electronic clinical data. He will discuss how there is an unprecedented opportunity to implement new measurements to identify risks and events related to the diagnostic process. He will use some of the audience responses to frame his comments.

Panel members represent a diversity of people, thought, expertise, and experience.

Topic Rationale
The session will focus on the intersection between informatics, patient safety and diagnostic accuracy and discuss implementation of recommendations and innovations to understand and improve diagnostic safety. It will thus be of interest to any informaticists with interest in using HIT to prevent diagnostic harm, a high-priority patient safety challenge for clinicians, health care organizations and policy stakeholders.

Over a period of about 45 minutes, moderated audience interactions will explore questions related to:

1. What are the most common measurement challenges encountered in using LHS approaches to improving diagnosis?
2. What types of experiences have audience members and their organizations had with overcoming measurement issues, including access to EHR data warehouses to develop algorithms to identify patients with quality and safety issues?
3. What types of methods and models are most useful in HIT-enabled measurement of diagnostic error from systems, providers and patients? How have audience members used EHR data in measurement of communication or diagnosis-related care gaps?
4. How can application of electronic health record (EHR)-based triggers be used to improve measurement of diagnostic safety? How do we encourage health systems and CMIOs to help perform measurement related to diagnostic safety in the absence of any incentives and in the presence of so many competing priorities?
5. How can a measurement approach that uses a specific set of diagnostic safety triggers be implemented in an organization using EHRs? Who would be the key stakeholders? How will data be analyzed and who would the data be communicated to? What sort of dashboards may be useful? How can better feedback systems to clinicians be better developed?
Participation statement
All proposed panelists are aware of this panel submission and have agreed to participate in the panel.

Panel Participant Details

- **Hardeep Singh, MD, MPH** is Chief of Health Policy, Quality & Informatics Program at the VA Center for Innovations in Quality, Effectiveness and Safety (IQuESt), a practicing internist at Michael E. DeBakey VA Medical Center and Professor of Medicine at Baylor College of Medicine.
- **Dean Sittig, PhD** is a professor of biomedical informatics at the University of Texas Health Science Center in Houston, TX. Dr. Sittig will be one of the discussants.
- **Ashley N.D. Meyer, PhD** is a cognitive psychologist and a VA funded health services researcher and career development awardee at IQuESt in Houston, TX.
- **Traber D. Giardina, PhD, MSW** is a patient safety researcher and an AHRQ funded K01 awardee. She is an Assistant Professor at Baylor College of Medicine in Houston, TX.
- **Divvy K. Upadhyay, MBBS, MPH** is the inaugural Researcher-in-Residence of the Safer Dx Learning Lab and scientist in the Division of Quality and Safety at Geisinger.

References

8. Meyer AND, Singh H. The path to diagnostic excellence includes feedback to calibrate how clinicians think *JAMA* 2019(Online First).

The learning objectives of this panel include:

1) how HIT-enabled health care organizations can become LHS for improving diagnosis,
2) How informaticists can help implement HIT-based pragmatic approaches to measure and understand diagnostic errors and engage in development of institutional infrastructure and feedback processes to promote individual and organizational learning.
**Dissemination of Tuberculosis Clinical Evidence - An Application of Network Analysis to Publicly Available Resources**

Meredith Abrams, MS, MPH; Dongwen Wang, PhD  
Biomedical Informatics, Arizona State University, Scottsdale, AZ

**Abstract**

Effective dissemination of Tuberculosis (TB) clinical evidence to healthcare providers is essential to address this pandemic. To identify, organize, and aggregate online TB information resources, we analyzed the websites of four CDC-funded TB Centers of Excellence (COE), identified the hosted resources, examined the outward linkages, and collected the external resources. We obtained 154 primary resources from TB COEs and 1521 linkages to external resources. We leveraged a network analysis approach to construct resource networks at the individual resource and parent site levels. We computed the degree of connectivity and PageRank for each node in the networks. Based on these analyses, we successfully identified important TB information hosting sites beyond the four TB COEs and individual TB resources widely cited. The initial results indicate that it is promising to use network analysis approaches to identify online clinical information resources. The generalizability of this approach needs to be examined in future research.

1. **Introduction**

One-fourth of the world’s population is estimated to be infected with Tuberculosis (TB)\(^1\). Quick dissemination of TB clinical research is imperative to minimize the impacts of this disease. TB testing and treatment recommendations are often complicated due to the nature of the disease and the dependence on a patient’s pre-existing medical conditions. Clinical guidelines for TB can become very long, complex, and arduous for healthcare providers to understand, to implement, and to adhere to\(^2\). To address these challenges, various online resources and tools have been created to facilitate TB clinical education and patient care through clinical decision support.

The Centers for Disease Control and Prevention (CDC) is the prevailing authority in the United States for development and dissemination of clinical evidence for public health. CDC has funded four regional Centers of Excellence (COE) for TB Training, Education, and Medical Consultation: (1) Curry International Tuberculosis Center (CITC); (2) the Global Tuberculosis Institute (GTI); (3) Heartland National Tuberculosis Center (HNTC); and (4) Southeastern National Tuberculosis Center (SNTC)\(^3\). The missions of these TB COEs are: “increasing human resource development through education and training activities” and “increasing the capacity for appropriate medical evaluation and management of persons with TB disease and latent TB infection through medical consultation”\(^4\). To serve for these missions, the four COEs have accumulated a variety of online TB information resources. Many of these online information resources are inter-connected through hyperlinks (for webpages) or citations (for documents).

Analyzing the linkages among the existing internet-based resources is a hot research topic in computer science. A popular research method is network analysis, which “can help in identifying the underlying network of relationships in a collaboration that people rely on to find information and solve problems”\(^5\). Network analysis assists in delineating the underlying structures within organizations or, in the case of online information, identifying effective approaches for resource dissemination\(^6\). It can also be utilized as “an invaluable tool for visualizing the structure of a collaboration and in identifying key roles, information flows and collaboration”\(^6\). Yet, few studies have leveraged network analysis to determine potential key sites of online information resources for specific clinical domains or diseases\(^6\).

The primary aim of this research is to collect, organize and aggregate online TB information resources. The secondary aim of this study is to identify the critical online TB information resources and their hosting websites through network analysis. The overarching goal of our work is to develop effective approaches to disseminate TB clinical evidence.

2. **Methods**

2.1 **Collection of Online TB Resources**

To identify the available online TB resources, we reviewed the websites of the four regional TB COEs for publicly available resources. Each TB COE maintains on its website a resource section that contains a list of materials for healthcare providers to utilize in practice.
2.2 Identification of External Resources through Outward Linkages

We reviewed the “outward linkages” from each TB COE’s website to identify the linked resources. For this purpose, a co-author (MA) evaluated each page from the websites of the four TB COEs and assessed all outward hyperlinks from the webpages and references from the documents to identify the external resources.

2.3 Analyses of TB Resources and Linkages

We analyzed the identified TB resources and linkages with respect to the following characteristics:

1) Format of Resource
We analyzed the format of the linked resources and identified the specific types of resources such as webpages, pdf documents, and multimedia objects (for example, videos).

2) Content of Resource
We reviewed and recorded the specific content of each linked resource to identify the detailed information provided.

3) Category/Theme of Resource
We catalogued the linked resources into specific themes. These themes were derived from a detailed review of the content of each target resource. For example, the theme “medication administration” includes all resources that contain information about medication injections or specific antibiotic instructions.

4) Linkage Context
We assessed the context for each outward linkage to determine the contextual clues of a linkage from a hosting webpage to a target resource.

5) Parent Site of Target Resource
For each target resource, we recorded the information of its parent website. This analysis aimed to identify the organizations that sponsor the target resources.

2.4 Construction and Analysis of TB Resource Network

We constructed the TB resource networks on two separate levels:

1) Individual Resource Level
We treated each linked resource as a node. We treated a linkage (hyperlink or citation) from a resource A to another resource B (webpage, document, or multimedia object) as a directed edge from node A to node B. Each referral to a target resource counted as a single instance. If a resource A is linked to the same target resource B multiple times, the total number of the linkages was assigned as the weight of the directed edge from A to B.

2) Parent Site Level
We treated each parent site as a node. We treated a linkage (hyperlink or citation) from any resource on parent site A to any target resource on parent site B as a directed edge from node A to node B. If a parent site A is linked to a parent site B multiple times, the total number of the linkages was assigned as the weight of the directed edge from A to B.

With the TB resource network constructed, we computed the following measures at both the individual resource and parent site levels for network analyses:

1) Degree of Connectivity
At the individual resource level, we calculated the degrees of connectivity to identify the number of inward and outward linkages for each linked resource. The degree of connectivity is defined as the “the number of nodes that a focal node is connected to, and measures the involvement of the node in the network”7. For this study, out-degree was calculated as the number of outward linkages from a specific resource, whereas in-degree was calculated as the number of inward linkages to that resource. The total degree was calculated as the sum of the in- and out-degrees.
At the parent site level, we treated each parent site as a node and calculated the in-degree and out-degree based on the linkages between the sites. Since we identified all the resources through the outward linkages from the four TB COEs (see Sections 2.1 and 2.2), only those four TB COE sites have an out-degree. Meanwhile, all the identified resources have an in-degree. We therefore focused on in-degree for the analysis at the parent site level.

2) PageRank

The PageRank measure was originally developed by Google as a means to indicate the “importance” of a single webpage. In network analysis, “PageRank can be thought of as a model of user behavior”8. This model identifies the probability that a random user transitions from one page to another and thus gives a score of “importance” calculated as a probability to each page in the network.

The calculation of PageRank for an example page “A” is based on the following formula:

\[ PR(A) = (1-d) + d \left( \frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)} \right) \]

Here \( T_1 \) to \( T_n \) are all the other webpages with hyperlinks pointing to \( A \), \( C(T_i) \) is the count of hyperlinks from webpage \( T_i \) to \( A \), where \( d \) is defined as a damping factor, set between 0 and 1, to indicate the probability a user continues to browse webpages following a hyperlink. A common default value for \( d \) is 0.85, which was used for this study8. We utilized a software package for network analysis, Gephi (version 0.9.2), to calculate the PageRank for each individual resource and parent site.

2.5 Visualization

Visualization is an intuitive approach to allow direct observation of the structure and other characteristics of a network. For this study, we leveraged the Gephi software package for visualizations.

3. Results

3.1 Online Resources Published by the Four TB COEs

Over the period from February 2018 through December 2018, we analyzed a total of 154 online resources published by the four TB COEs. A majority of these resources (133, 86.4%) were downloadable PDF documents or informational webpages. The remaining 21 (13.6%) resources included multimedia files, toolkits, and continuing medical education (CME) courses. With regard to the topics and prevailing themes of the resources, we identified 29 (18.8%) patient education materials, 19 (12.3%) geographic guidance for clinicians, and 13 (8.4%) for cultural competence training. Among the identified resources, 123 (79.9%) included the date for original publication or revision, ranging from 2001 to 2017. In terms of the size of the materials, the PDF documents ranged from 1 to 326 pages (on average 33 pages), and the multimedia files ranged from 31 seconds to 85 minutes (on average 28 minutes and 29 seconds). A summary of the identified TB resources and their catalogue is shown in Table 1.

3.2 Identification of Outward Linkages

We recorded a total of 1521 outward linkages from the 154 resources identified. The top linkage contexts included 494 citations (32.48%), 359 referrals for detailed information on specific topics (23.60%), 307 links to patient appropriate educational resources (20.18%), 270 general resources such as printable documents or videos (17.75%), and 29 provider education materials (1.91%). The top linkage topics included materials related to improving clinician’s cultural competency (521, 36.28%), general TB information (254, 17.69%), TB/HIV coinfection (112, 7.80%), TB medication (88, 6.13%), and TB directly observed therapy (84, 5.85%). The parent sites with the most linked target resources included CDC (8.42%), CITC (5.98%), World Health Organization (WHO) (5.92%), SNTC (3.88%), GTI (3.74%), EthnoMed9 (2.43%), and the United States Department of Health Services (USDHS) (1.97%). The details of the outward linkage metrics are shown in Table 2.

3.3 Network Analysis - Degree of Connectivity

At the parent site level, we recorded an out-degree of 170 for SNTC, 113 for CITC, 38 for GTI, and 30 for HNTC. The in-degree of the four TB COEs is between 3 and 4, and the total degree ranges from 33 to 174. The details of this analysis are shown in Table 3.
Table 1. Catalogue of TB Resources.

<table>
<thead>
<tr>
<th>Hosting TB COE</th>
<th>N</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CITC</td>
<td>25</td>
<td>16.23</td>
</tr>
<tr>
<td>GTI</td>
<td>34</td>
<td>22.07</td>
</tr>
<tr>
<td>HNTC</td>
<td>32</td>
<td>20.77</td>
</tr>
<tr>
<td>SNTC</td>
<td>63</td>
<td>40.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Formats</th>
<th>N</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static PDF</td>
<td>117</td>
<td>75.97</td>
</tr>
<tr>
<td>Informational Webpage</td>
<td>16</td>
<td>10.38</td>
</tr>
<tr>
<td>Video/multimedia presentation</td>
<td>11</td>
<td>7.14</td>
</tr>
<tr>
<td>Toolkits</td>
<td>7</td>
<td>4.54</td>
</tr>
<tr>
<td>Downloadable Audio File</td>
<td>2</td>
<td>1.29</td>
</tr>
<tr>
<td>CEI Credit Course</td>
<td>1</td>
<td>0.65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
<th>N</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Printable Material: Single page</td>
<td>14</td>
<td>9.09</td>
</tr>
<tr>
<td>Printable Material: 2-10 pages</td>
<td>40</td>
<td>25.97</td>
</tr>
<tr>
<td>Printable Material: 11-50 pages</td>
<td>40</td>
<td>25.97</td>
</tr>
<tr>
<td>Printable Material: 51-100 pages</td>
<td>9</td>
<td>5.84</td>
</tr>
<tr>
<td>Printable Material: 101 pages and above</td>
<td>14</td>
<td>9.09</td>
</tr>
<tr>
<td>Multimedia File: 0-15 minutes</td>
<td>4</td>
<td>2.59</td>
</tr>
<tr>
<td>Multimedia File: 15-60 minutes</td>
<td>6</td>
<td>3.89</td>
</tr>
<tr>
<td>Multimedia File: Greater than 60 min</td>
<td>1</td>
<td>0.65</td>
</tr>
<tr>
<td>Not Applicable</td>
<td>26</td>
<td>16.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prevailing Theme</th>
<th>N</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient education</td>
<td>29</td>
<td>18.83</td>
</tr>
<tr>
<td>Region-Specific Guidance</td>
<td>19</td>
<td>12.34</td>
</tr>
<tr>
<td>Cultural Competence</td>
<td>13</td>
<td>8.44</td>
</tr>
<tr>
<td>TB Infection Control</td>
<td>12</td>
<td>7.79</td>
</tr>
<tr>
<td>Provider Education</td>
<td>11</td>
<td>7.14</td>
</tr>
<tr>
<td>TB Treatment</td>
<td>10</td>
<td>6.49</td>
</tr>
<tr>
<td>Harm reduction information</td>
<td>9</td>
<td>5.84</td>
</tr>
<tr>
<td>Radiographic Information</td>
<td>9</td>
<td>5.84</td>
</tr>
<tr>
<td>Case Management</td>
<td>8</td>
<td>5.19</td>
</tr>
<tr>
<td>Contact Investigation Skills</td>
<td>7</td>
<td>4.55</td>
</tr>
<tr>
<td>Archived Guideline</td>
<td>6</td>
<td>3.89</td>
</tr>
<tr>
<td>Medication Administration</td>
<td>4</td>
<td>2.60</td>
</tr>
<tr>
<td>Pediatric TB</td>
<td>3</td>
<td>1.95</td>
</tr>
<tr>
<td>TB Evaluation</td>
<td>3</td>
<td>1.95</td>
</tr>
<tr>
<td>Nutrition Recommendations</td>
<td>2</td>
<td>1.30</td>
</tr>
<tr>
<td>Testing resources</td>
<td>2</td>
<td>1.30</td>
</tr>
<tr>
<td>Teaching Curricula</td>
<td>2</td>
<td>1.30</td>
</tr>
<tr>
<td>Ethics</td>
<td>2</td>
<td>1.3</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>3</td>
<td>1.95</td>
</tr>
</tbody>
</table>

After combined all the resources from the four TB COEs and other organizations, the top parent sites based on the in-degree are: SNTC, GCI, CDC, WHO, Management Sciences for Health, HNTC, CITC, University of California San Francisco, EthnoMed, Find TB Resources, TB Controllers of America, Diversity Rx, and Stop TB Partnership. The remaining parent sites had an in-degree below 3. The detailed result of this analysis is shown in Table 4.
Table 2. Outward Linkage Metrics

<table>
<thead>
<tr>
<th>Top Linkage Contexts</th>
<th>N</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations</td>
<td>494</td>
<td>32.48</td>
</tr>
<tr>
<td>Referral for Specific Information</td>
<td>359</td>
<td>23.60</td>
</tr>
<tr>
<td>Referral Patient Education Resources</td>
<td>307</td>
<td>20.18</td>
</tr>
<tr>
<td>Referral for General Resources (printable/video)</td>
<td>270</td>
<td>17.75</td>
</tr>
<tr>
<td>Provider Education Materials</td>
<td>29</td>
<td>1.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top Linkage Topics</th>
<th>N</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cultural Competency</td>
<td>521</td>
<td>36.28</td>
</tr>
<tr>
<td>General TB information</td>
<td>254</td>
<td>17.69</td>
</tr>
<tr>
<td>TB/HIV Coinfection</td>
<td>112</td>
<td>7.80</td>
</tr>
<tr>
<td>TB Medication</td>
<td>88</td>
<td>6.13</td>
</tr>
<tr>
<td>TB Directly Observed Therapy</td>
<td>84</td>
<td>5.85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top Parent Sites</th>
<th>N</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC</td>
<td>128</td>
<td>8.42</td>
</tr>
<tr>
<td>CITC*</td>
<td>91</td>
<td>5.98</td>
</tr>
<tr>
<td>WHO</td>
<td>90</td>
<td>5.92</td>
</tr>
<tr>
<td>SNTC*</td>
<td>59</td>
<td>3.88</td>
</tr>
<tr>
<td>GTI*</td>
<td>57</td>
<td>3.74</td>
</tr>
<tr>
<td>EthnoMed*</td>
<td>37</td>
<td>2.43</td>
</tr>
<tr>
<td>USDHS</td>
<td>30</td>
<td>1.97</td>
</tr>
</tbody>
</table>

* TB COE

Table 3. Degree of Connectivity for the Four TB COEs

<table>
<thead>
<tr>
<th>Parent Site Organization</th>
<th>Out-Degree</th>
<th>In-Degree</th>
<th>Total Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNTC</td>
<td>170</td>
<td>4</td>
<td>174</td>
</tr>
<tr>
<td>CITC</td>
<td>113</td>
<td>3</td>
<td>116</td>
</tr>
<tr>
<td>GTI</td>
<td>38</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>HNTC</td>
<td>30</td>
<td>3</td>
<td>33</td>
</tr>
</tbody>
</table>

Table 4. Top Parent Sites Based on the In-Degree

<table>
<thead>
<tr>
<th>Parent Site Organization</th>
<th>In-Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNTC*</td>
<td>4</td>
</tr>
<tr>
<td>GTI*</td>
<td>4</td>
</tr>
<tr>
<td>CDC</td>
<td>4</td>
</tr>
<tr>
<td>WHO</td>
<td>4</td>
</tr>
<tr>
<td>Management Sciences for Health</td>
<td>4</td>
</tr>
<tr>
<td>HNTC*</td>
<td>3</td>
</tr>
<tr>
<td>CITC*</td>
<td>3</td>
</tr>
<tr>
<td>University of California at San Francisco</td>
<td>3</td>
</tr>
<tr>
<td>EthnoMed*</td>
<td>3</td>
</tr>
<tr>
<td>Find TB Resources</td>
<td>3</td>
</tr>
<tr>
<td>TB Controllers of America</td>
<td>3</td>
</tr>
<tr>
<td>Diversity Rx*</td>
<td>3</td>
</tr>
<tr>
<td>Stop TB Organization*</td>
<td>3</td>
</tr>
</tbody>
</table>

* TB COE
At the individual resource level, the resources with the highest in-degree included: SNTC Main Products page (17), TinyURL Webhosting Page for TB Fotonovela (8), HNTC homepage (8), SNTC homepage (8), New Mexico Department of Health Active TB informational page (8), New Mexico Department of Health Sputum Collection Instruction page (8), US Department of Homeland Security 2008 report on Profiles of Naturalized Citizens for the fiscal year of 2006 (8), GTI homepage (7), and US Census Bureau Fact Finder page (7). The detailed result of this analysis is shown in Table 5.

Table 5. Top Individual Resources Based on the In-Degree

<table>
<thead>
<tr>
<th>Individual Resource</th>
<th>Parent Organization</th>
<th>In-Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNTC Main Products Page</td>
<td>SNTC*</td>
<td>17</td>
</tr>
<tr>
<td>TinyURL Webhosting Page for TB Fotonovela</td>
<td>TinyURL Webhost</td>
<td>8</td>
</tr>
<tr>
<td>HNTC Homepage</td>
<td>HNTC*</td>
<td>8</td>
</tr>
<tr>
<td>SNTC Homepage</td>
<td>SNTC*</td>
<td>8</td>
</tr>
<tr>
<td>Active TB Informational Page</td>
<td>New Mexico DOH</td>
<td>8</td>
</tr>
<tr>
<td>Sputum Collection Instruction Webpage</td>
<td>New Mexico DOH</td>
<td>8</td>
</tr>
<tr>
<td>Report on Profiles of Naturalized Citizens for the Fiscal Year 2006</td>
<td>USDHS</td>
<td>8</td>
</tr>
<tr>
<td>GTI Homepage</td>
<td>GTI*</td>
<td>7</td>
</tr>
<tr>
<td>US Census Bureau Fact Finder Page</td>
<td>US Census</td>
<td>7</td>
</tr>
</tbody>
</table>

*TB COE

3.4 Network Analysis – PageRank

At the parent site level, the organizations with the top PageRanks were: CDC (0.003654), WHO (0.003654), Management Science for Health (0.003654), SNTC (0.003636), TB Controllers of America (0.003636), Diversity Rx10 (0.003636), University of California at San Francisco (0.003627), CITC (0.003609), GTI (0.003574), and Stop TB Partnership11 (0.003574), as shown in Table 6.

Table 6. Top 10 Parent Sites by PageRank

<table>
<thead>
<tr>
<th>Parent Site Organization</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC</td>
<td>0.003654</td>
</tr>
<tr>
<td>WHO</td>
<td>0.003654</td>
</tr>
<tr>
<td>Management Sciences for Health</td>
<td>0.003654</td>
</tr>
<tr>
<td>SNTC*</td>
<td>0.003636</td>
</tr>
<tr>
<td>TB Controllers of America</td>
<td>0.003636</td>
</tr>
<tr>
<td>Diversity Rx10</td>
<td>0.003636</td>
</tr>
<tr>
<td>University of California at San Francisco</td>
<td>0.003627</td>
</tr>
<tr>
<td>CITC*</td>
<td>0.003609</td>
</tr>
<tr>
<td>GTI*</td>
<td>0.003574</td>
</tr>
<tr>
<td>Stop TB Partnership11</td>
<td>0.003574</td>
</tr>
</tbody>
</table>

* TB COE

At the individual resource level, the webpages with the top PageRanks were: SNTC Main Products page (0.008372), TinyURL Webhosting Page for TB Fotonovela (0.003547), GTI homepage (0.003262), GTI Resources page (0.002305), HNTC Medical Consultation page (0.002251), Drugs.com Amikacin Administration information page (0.002134), Nursing Center Take 5 Z-Track Injection informational page (0.002134), CITC Medication Fact Sheet page (0.001494), GTI TB Interviewing for Contact Investigation: A Practical Resource for the Healthcare Worker page (0.001488), and WHO Regional Office for Europe homepage (0.001488). It is important to note that 6 out of the top 10 resources were hosted by the four TB COEs. The details of the results can be found in Table 7.
### Table 7. Top Individual Resources by PageRank

<table>
<thead>
<tr>
<th>Webpage</th>
<th>Parent Organization</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNTC Main Products Page(^\text{12})</td>
<td>SNTC*</td>
<td>0.008372</td>
</tr>
<tr>
<td>Tiny URL Host for TB Fotonovela(^\text{13})</td>
<td>TinyURL Webhost</td>
<td>0.003547</td>
</tr>
<tr>
<td>GTI Homepage(^\text{14})</td>
<td>GTI*</td>
<td>0.003262</td>
</tr>
<tr>
<td>GTI Resources Page(^\text{15})</td>
<td>GTI*</td>
<td>0.002305</td>
</tr>
<tr>
<td>HNTC Medical Consultation Page(^\text{16})</td>
<td>HNTC*</td>
<td>0.002251</td>
</tr>
<tr>
<td>Amikacin Administration Information Page(^\text{17})</td>
<td>Drugs.com</td>
<td>0.002134</td>
</tr>
<tr>
<td>Take 5 Z-Track Injections Information Page(^\text{18})</td>
<td>Nursing Center</td>
<td>0.002134</td>
</tr>
<tr>
<td>Medication Fact Sheet Page(^\text{19})</td>
<td>CTC*</td>
<td>0.001494</td>
</tr>
<tr>
<td>TB Interviewing for Contact Investigation: A Practical Resource for the Healthcare Worker(^\text{20})</td>
<td>GTI*</td>
<td>0.001488</td>
</tr>
<tr>
<td>WHO Regional Office for Europe Homepage(^\text{21})</td>
<td>WHO</td>
<td>0.001488</td>
</tr>
</tbody>
</table>

* TB COE

#### 3.5 Visualizations

Visualization of the TB resource network based on the PageRank of parent site is shown in Figure 1. A parent site with a higher PageRank is presented as a node with a larger size. The thickness of an edge indicates the weight of that edge. The top 10 parent sites and all four TB COEs are pinpointed in Figure 1.

Visualization of the TB resource network based on the PageRank of individual resource is shown in Figure 2. Resources hosted under a top-ranked parent site are presented as nodes in the same color. A resource with a higher PageRank is presented as a node with a larger size. The thickness of an edge indicates the weight of that edge. The details of a specific top-ranked individual resource is zoomed at the lower-right corner of Figure 2.

![Figure 1. Visualization of the TB Resource Network Based on the PageRank of Parent Websites](image-url)
4. Discussion

In this study, we assessed the existing web-based TB resources and their linkages. For this purpose, we started from the websites of the four national TB COEs and identified additional resources through hyperlinks and citations. We examined each resource to record its format, content, and theme. This analysis provided important information to understand the structure of online resources, the characteristics of these resources, and the potential strategies for their dissemination, which are essential for biomedical informatics research.

By leveraging network analysis approaches and the associated software tools, we constructed TB online resource networks at both the individual resource and the parent site levels. In particular, we identified major TB resource hosting sites beyond the four TB COEs, such as CDC, WHO, Management Science for Health, University of California at San Francisco, TB Controllers of America, Diversity Rx, and Stop TB Partnership. While sites like CDC and WHO are well known for their informational resources, this analysis identified less-known websites, such as EthnoMed and Diversity Rx, which hosted highly useful information for TB healthcare providers.

We utilized two network measures to identify the top sites, i.e., degree of connectivity and PageRank. Each measure has its own advantages and limitations. The degree of connectivity only focuses on the local structure, which makes it very easy to compute. However, simply linked to other nodes may not necessarily mean that such linkages are important. PageRank provides a quantitative measure on value of different linkages by considering the global importance of the linked nodes, and therefore can provide a more accurate evaluation of resources. On the other hand, calculation of PageRank becomes more complex due to this global consideration. It is interesting to note that a significant portion of the identified top sites based on degree of connectivity are overlapped with those based on PageRank. A potential explanation is that we only had two levels of resource linkages when developing the network (the four TB COEs and their linked resources) and therefore the difference between the local and global structures is...
insignificant. Our future work will explore multiple levels of resource linkages and re-examine the use of specific network measures to identify important hosting sites for TB resources.

Through analysis at the individual resource level, we identified several useful TB resources, such as those about sputum collection and active TB hosted at New Mexico Department of Health. These two resources had a high level of connectivity, with linkages from all four TB COEs. We also found external resources with a high PageRank, such as the Amikacin Administration page hosted at Drugs.com. By identifying resources with high PageRank, we can further examine the specific information that is perceived as important by the TB COEs and their targeting clinicians.

Visualization of the TB resource networks allowed us to directly examine the important parent sites or individual resources identified from the analysis. For example, visualization of the parent site network placed CDC and WHO at the central location, with a node size (indicating PageRank) even larger than the four TB COEs. Visualization of the individual TB resources, on the other hand, requires more efforts to examine due to the large number of nodes in the network. Nevertheless, it still allowed us to identify certain important resources, such as the GTI homepage. Meanwhile, the color coding of resources made it intuitive to observe those hosted at the same parent site as well as the linkages within and beyond the hosting site.

There are a few limitations in this research. First, the collection of online TB information resources was constrained to only the four TB COEs and their direct linkages. Although these four TB COEs represent the major national resources for TB, the completeness of the collected TB resources needs to be further verified. Second, all the online resources were manually reviewed by a single person, which may introduce potential selection bias and reduce the reliability of data collection. Although this reviewer has extensive experience working with TB, ideally we should have multiple independent reviewers to ensure a high quality and fair process of the analyses. Third, the network analysis was limited to two levels of resource linkages. To further examine the use of specific network measures to identify the important resources and the hosting sites, we will need to perform network analysis with multiple levels of resources linkages. Last, this study focused specifically on TB information resources and their linkages. The generalizability of the findings from this study to other diseases and clinical problems needs to be verified in future studies.

Our work in the next steps includes: (1) expanding the review of TB resources to the sites beyond the four COEs; (2) leveraging web crawling technologies to automatically identify linked resources and to facilitate the construction of resource networks; (3) integrating semantic analysis techniques to provide another perspective on content of the resources identified; (4) performing larger scale network analyses with multi-level resource linkages and additional measurement on network properties to identify a complete set of TB information resources; and (5) generalizing the study approaches for applications to other diseases, clinical problems, and public health challenges.

5. Conclusion

We have analyzed online TB information resources through manual reviewing the websites of the four national TB COEs and their linkages. Based on this analysis, we have successfully constructed TB resource networks at both the parent site and the individual resource levels. Leveraging network analysis methods and the associated tools, we have identified important TB resources and hosting sites. The initial results from this study indicates that it is promising to use network analysis approaches for identification and dissemination of online clinical information resources. Our future work will expand the scope of resource review, develop automated processes to identify resources and linkages, perform larger scale network analysis, and examine the generalizability of the study approaches, with an ultimate goal for more effective dissemination of clinical evidence and better care of patients with TB and other diseases.

References


Evaluating the Portability of an NLP System for Processing Echocardiograms: A Retrospective, Multi-site Observational Study

Prakash Adekkanattu, PhD1, Guoqian Jiang, MD PhD2, Yuan Luo, PhD3, Paul R. Kingsbury, PhD2, Zhenxing Xu, PhD1, Luke V. Rasmussen, MS3, Jennifer A. Pacheco, MS3, Richard C. Kiefer, MS2, Daniel J. Stone, MS3, Pascal S. Brandt, MS4, Liang Yao, PhD3, Yizhen Zhong, BS3, Yu Deng, BS3, Fei Wang, PhD1, Jessica S. Ancker, MPH PhD1, Thomas R. Campion, Jr, PhD1, Jyotishman Pathak, PhD1

1Weill Cornell Medicine, New York, NY, 2Mayo Clinic, Rochester, MN, 3Northwestern University, Chicago, IL, 4University of Washington, Seattle, WA

Abstract

While natural language processing (NLP) of unstructured clinical narratives holds the potential for patient care and clinical research, portability of NLP approaches across multiple sites remains a major challenge. This study investigated the portability of an NLP system developed initially at the Department of Veterans Affairs (VA) to extract 27 key cardiac concepts from free-text or semi-structured echocardiograms from three academic medical centers: Weill Cornell Medicine, Mayo Clinic and Northwestern Medicine. While the NLP system showed high precision and recall measurements for four target concepts (aortic valve regurgitation, left atrium size at end systole, mitral valve regurgitation, tricuspid valve regurgitation) across all sites, we found moderate or poor results for the remaining concepts and the NLP system performance varied between individual sites.

Introduction

Echocardiography is currently the most widely used non-invasive cardiac imaging in the diagnosis, treatment, and follow-up of patients with cardiovascular diseases. Various measurements in echocardiography provide key insights into the mechanisms and health of a patient’s cardiovascular system. Aortic, mitral and tricuspid valves are assessed for abnormal structure, thickness, and valve dysfunction (regurgitation, stenosis). Left ventricular systolic performance indicates severity of heart disease, and most laboratories quantify Left Ventricular Ejection Fraction (LVEF) to a value ranging from 5-75 percent of total blood volume. Echocardiography measurements are generally stored as an echocardiogram (echo) report in electronic health record (EHR) systems, which is frequently an unstructured or semi-structured text document. This limits the downstream use of the report data by automated systems for clinical research (e.g., phenotyping), care management and point-of-care clinical decision support.

Natural language processing (NLP) of unstructured clinical narratives, such as echo reports, holds tremendous potential to extract meaningful information and fill this knowledge gap. NLP utilizes various algorithms to automatically extract relevant clinical information from free text and semi-structured data sources. However, a recent review of the literature by Demner-Fushman and Elhadad suggests that NLP remains an “emerging technology”, with a significant gap between promise and reality. One of the major challenges in broader adoption of NLP systems and technologies is their portability across multiple EHR systems. While several studies, including work done by our study team, have demonstrated varying levels of success in portability of NLP technologies across institutions, recent work by Carrell et al. argue that there remain significant challenges in adapting NLP systems across multiple sites, which include assembling clinical corpora, managing diverse document structures and handling idiosyncratic linguistic expressions. These barriers and the requirement for significant upfront “investment” limit wide-scale adoption of NLP tools and systems across health systems, particularly those in low-resource settings and environments.

To understand these barriers further, in this study we investigated the implementation of an existing NLP system - Leo, developed by the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI) - across three academic medical centers. The goal of the current study was to assess the portability of EchoExtractor by installing, without any system modifications whatsoever, and evaluating its performance at three non-VA medical centers. This project, as part of the Phenotype Execution and Modeling Architecture (PhEMA), was motivated by the fact that EHR based computational phenotyping using both structured and unstructured data is a key requirement for deriving high quality phenotypes from EHRs, and a portable NLP system that can be run across multiple institutions is much needed.
Materials and Methods

System Description and Target Concepts

Leo provides a set of services and libraries that facilitate the rapid creation and deployment of NLP analysis tools. A major strength of this approach is the use of Unstructured Information Management Architecture (UIMA), which provides a set of standards for creating, discovering, composing and deploying a broad range of text analytics capabilities and integrating them with search technologies. This standard is very attractive for NLP research because it is open-source (available through Apache), and enables incorporation of NLP tools developed elsewhere. The distribution package for EchoExtractor was made available through a VA GitHub repository. Installation of EchoExtractor using UIMA-AS 2.8.1 was deemed to be relatively straightforward, and was completed within a day at WCM and Northwestern. Installation at Mayo Clinic was performed using installation instructions found for Leo where an alternate version of UIMA (UIMA-AS 2.9.0) was installed to support Leo, yet the installation process was still also completed within a day. We tested the performance of the pipeline without any custom modification to evaluate the true portability of the NLP system.

While previous studies have shown success in extracting echocardiography measurements from Echo reports using NLP methods, the focus has been primarily on extracting one key concept - LVEF. The VINCI team created a specific instance of Leo named EchoExtractor to extract 27 different important cardiac concepts from echo and radiology reports. A preliminary analysis of results obtained after processing 100 echo reports from WCM center through the NLP pipeline revealed that not all target concepts were present. Concepts such as left ventricular contractility, left ventricular hypertrophy and aortic valve max pressure gradient were not identified. Therefore, we focused on the remaining 24 elements in our extended evaluation study. Later, we observed that the above three concepts were in fact present in echo reports originated from other centers, however we did not include them in our performance evaluation. The 24 target concepts we investigated in this study is listed in Table 1. For a description of these concepts, we refer readers to Patterson, et al.

Data collection

We obtained Echo reports from all three medical centers participating in this study. Weill Cornell Medicine (WCM) uses the EpicCare® Ambulatory EHR platform to document clinical care, from which we collected all echocardiograms used as the basis of this study. We obtained 200 echocardiograms for 200 patients who were 18 years or older at the time of the echocardiogram results and one of the study team members (PA) manually annotated the data elements extracted by the EchoExtractor to create the gold standard at WCM. These reports in general formatted into multiple sections containing semi-structured, and unstructured data. Reports have a subject meta-data section, one or more tabular sections, and a comment section. The tabular sections contain quantitative measures such as wall thicknesses, chamber dimensions, or flow velocities. Unstructured fields contain descriptions of clinically relevant findings as interpreted by the technician. For example, the descriptions could say “Normal aortic valve and aortic root”, “Normal left ventricular size and function”, “There is trace mitral regurgitation”. Furthermore, left ventricular function is often subjectively quantified as “mild”, “moderate”, “mildly to moderately reduced” or “severely reduced”.

Mayo Clinics implemented an Unified Data Platform (UDP) to provide practical data solutions that creates a combined view of multiple heterogeneous EHR data sources (including Epic) through effective data orchestration, along with a number of data marts based on common data models. We retrieved the echocardiogram reports from the Mayo Clinic UDP platform from patients over 18 years old at the time of the echocardiogram results after verifying authorization to use the data in accordance with the Minnesota Health Records Act. Patient records were found in two formats: Rich Text Format (2000 to 2012), and Portable Document Format (PDF) from 2012 to 2018. For the echocardiograms generated between 2012-18, using a custom-built Python application, original text from the PDF documents was extracted. We randomly selected 200 echocardiogram reports for this study from each document format. The text portions of these reports were extracted, PHI was censored, and other elements such as visual elements (charts and graphics) and metadata were discarded. Echo reports found at Mayo Clinic generally consisted of three types of sections: unstructured text, numbered final impressions, and semi-structured tables of echocardiogram data. These reports were manually annotated with data elements extracted by the EchoExtractor by an experienced nurse at the Mayo Clinic which generated an annotation corpus as Mayo Clinic’s gold standard.
<table>
<thead>
<tr>
<th>Concept</th>
<th>WCM</th>
<th>MAYO</th>
<th>NU</th>
<th>MIMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>aortic valve mean gradient</td>
<td>P</td>
<td>A</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>aortic valve orifice area</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>aortic valve regurgitation</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>aortic valve regurgitation peak velocity</td>
<td>A</td>
<td>A</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>aortic valve stenosis</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>e/e prime ratio</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>inter-ventricular septum</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>dimension at end diastole</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>left atrium size at end systole</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>left ventricular dimension at end diastole</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>left ventricular dimension at end systole</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>left ventricular size</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>left ventricular ejection fraction</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>left ventricular posterior wall thickness at end diastole</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>mitral valve mean gradient</td>
<td>A</td>
<td>P</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>mitral valve orifice area</td>
<td>P</td>
<td>A</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>mitral valve regurgitation</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>mitral valve regurgitation peak velocity</td>
<td>A</td>
<td>A</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>mitral valve stenosis</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>pulmonary artery pressure</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>right atrial pressure</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>tricuspid valve mean gradient</td>
<td>A</td>
<td>A</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>tricuspid valve orifice area</td>
<td>A</td>
<td>A</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>tricuspid valve regurgitation</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>tricuspid valve regurgitation peak velocity</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
</tbody>
</table>

At Northwestern Medicine, we retrieved 251,521 free-text echocardiogram reports for patients over 18 years old at the time of the echocardiogram between 2000 and 2017 from Northwestern Medicine® Enterprise Data Warehouse (NMEDW). The NMEDW hosts a comprehensive and integrated repository of the most used clinical and research data sources, at the time from 2 hospitals: Northwestern Memorial Hospital (NMH) and Lake Forest Hospital (LFH). Echo reports typically contained a mixture of semi-structured tables of echocardiogram data, along with unstructured impressions and findings. Overall structure of reports was primarily a metadata section containing patient and procedure information, semi-structured tabular results, followed by the impression (although many variations in this structure were observed). We randomly selected 200 echocardiogram reports and multiple study team members (YZ, YD, LY, JP, LR, YL) manually annotated the data elements extracted by the EchoExtractor. Across reviewers, we have discordant findings reconciled in a group setting to ensure consistent coding and interpretation, sometimes in
consultation with a cardiovascular physician colleague. The final consolidated list of annotations was used as the gold standard at Northwestern.

We also obtained echocardiograms from the Medical Information Mart for Intensive Care III (MIMIC-III) database. The latest version, MIMIC-III, contains de-identified patient records for >40,000 critical care patients admitted to Beth Israel Deaconess Medical Center between 2001 and 2012. We extracted 44,559 echo reports from 21,645 adult patients who were 18 years or older. We then randomly selected 200 echo reports for 200 unique patients and a study team member (PA) manually annotated the data elements extracted by the EchoExtractor to create the gold standard for the MIMIC dataset. Unlike the Echo reports from three medical centers, MIMIC reports were in free-text format with no tabular or semi-structured sections. In general, these reports have an interpretation section comprising subsections for findings and conclusions. Findings section has detailed descriptions about left atrium, right atrium, left ventricle, right ventricle, aorta, aortic valve, mitral valve, tricuspid valve, and pulmonic valve/pulmonic artery. MIMIC reports are for patients admitted to the intensive care unit; hence, patient characteristics are different from those patients treated at various outpatient settings in WCM, Mayo Clinic and Northwestern. At each medical center we limited the number of reports for manual review to 200 mainly due to limited resource availability, and also due to the fact that, in general, echo reports tend to follow same format(s) within each clinical setting. Figure 1 shows typical examples of echo report from the four EHR systems investigated in this study.

![Figure 1. Typical examples of echocardiogram originated from multiple EHR systems](image)

**Extraction methodology**

Patterson et al. has described in detail the logic for concept extraction employed in the present study. Various target concepts were extracted through a multi-stage process in a rule-based algorithm that involves identifying a concept mention in the text, identifying a value mention in the text, and linking appropriate concept and value mentions in a relationship. Term identification and term mapping was used to detect a concept mention in the text. Term identification was performed using regular expressions matching and term mapping was achieved by building a custom term lookup dictionary. Value identification was achieved through regular expressions that match both quantitative and qualitative values.

**WCM**

- Peak Aortic: 2.1 m/s ± 0.1 m/s
- AVI: 50 mm Hg
- Right Atrium: 1.3 m/s ± 0.1 m/s
- Left ventricle: 1.5 m/s ± 0.1 m/s
- Left atrium: 0.9 mm Hg
- Right ventricle: 0.3 m/s ± 0.1 m/s
- Pulmonic valve: 0.7 m/s ± 0.1 m/s
- Aortic valve: 0.4 m/s ± 0.1 m/s
- Tricuspid valve: 0.6 m/s ± 0.1 m/s
- Stroke volume: 55 mL
- EF: 55%
- SV: 55 ml

**MAYO**

- Left ventricular ejection fraction: 60%
- Left atrium: 1.5 cm
- Right atrium: 1.2 cm
- Left ventricle: 3.0 cm
- Right ventricle: 1.8 cm
- Aorta: 2.5 cm
- Mitral valve: 2.3 cm
- Tricuspid valve: 2.1 cm
- Pulmonic valve: 2.2 cm
- Aortic valve: 2.0 cm
- Stroke volume: 60 mL
- EF: 60%

**NU**

- Left ventricular ejection fraction: 55%
- Left atrium: 1.3 cm
- Right atrium: 1.1 cm
- Left ventricle: 2.8 cm
- Right ventricle: 1.5 cm
- Aorta: 2.0 cm
- Mitral valve: 2.2 cm
- Tricuspid valve: 2.1 cm
- Pulmonic valve: 2.0 cm
- Aortic valve: 2.0 cm
- Stroke volume: 60 mL
- EF: 60%

**MIMIC**

- Left ventricular ejection fraction: 55%
- Left atrium: 1.3 cm
- Right atrium: 1.1 cm
- Left ventricle: 2.8 cm
- Right ventricle: 1.5 cm
- Aorta: 2.0 cm
- Mitral valve: 2.2 cm
- Tricuspid valve: 2.1 cm
- Pulmonic valve: 2.0 cm
- Aortic valve: 2.0 cm
- Stroke volume: 60 mL
- EF: 60%
and qualitative values for concepts. Finally, a relationship between a concept and value was established through pattern matching, from a set of patterns identified for the clinical subdomain of echocardiograms.

Reference standard and performance evaluations

For all three study sites and for the MIMIC dataset, we determined performance of the NLP pipeline individually by developing independent gold standard datasets (see above). Results of EchoExtractor output were tabulated and compared against the gold standard through manual reviews. For each concept, a reviewer compared the output generated by EchoExtractor to each of the 200 reports. The reviewer identified all mentions of a concept and the associated quantitative or qualitative values. The reviewer also confirmed that the concepts not found by EchoExtractor in a given report were truly not mentioned in the respective report. The reviewer identified both numeric values as well as qualitative descriptions such as normal, mild, moderate, dilated, and severe. If a numeric value is documented as a range, then we identified both the values as the minimum and maximum. In reports containing multiple instances of a concept, we identified all instances. In cases where concept value is expressed using a greater than or less than symbol (e.g. LVEF >55), the reviewer simply identified the value ignoring the symbol. Some echo reports express uncertainty about certain concepts such as LVEF value using a question mark (e.g. LVEF ?55-70). In such cases, the reviewer extracted the value, ignoring the question mark.

Each report was classified as one of four possible cases: true positive (report had the concept present and EchoExtractor identified the “concept-value pair” and matched with the value given in the reference standard); false positive (report had the concept absent, but EchoExtractor produced a concept-value pair); true negative (report had the concept absent, and EchoExtractor did not find any concept-value pair); and false negative (report had the concept present, but EchoExtractor did not identify a concept-value pair, or the value extracted did not match with the corresponding value in the text). A value for a concept can be either a quantitative measure as in “systolic function (LVEF>55%)” for ejection fraction, or a qualitative assessment as in “no aortic regurgitation”. Note that we are not assessing if a given concept is present or absent in the report, rather whether the NLP method correctly identified the concept-value pairs documented in the report. For example, for a statement such as “There is no aortic valve stenosis”, if the NLP system extracted ‘no’ as the assessment for the concept ‘aortic valve stenosis’, we will consider this as a true positive case. For a given target concept, when there were multiple instances of concept-value pair extracted by EchoExtractor, we used the following heuristic measure to select a given instance of the concept in order to compare directly with the reference standard: we selected the last concept-value pair, normally in the conclusion part of the report (as the order of results and text in the reports was usually chronological). The total outcomes of the four possible cases were then used to calculate various statistical performance measures, namely, precision (positive predictive value), recall (sensitivity or true positive rate), and the F-score (the harmonic mean of recall and precision).

Results

Results of EchoExtractor validation performed on echo reports from WCM, Mayo, Northwestern and MIMIC datasets are presented in Table 2. Among all datasets aortic valve regurgitation, aortic valve stenosis, left ventricular size, LVEF, mitral valve regurgitation, pulmonary artery pressure, and tricuspid valve regurgitation were the most frequently mentioned concepts. Other concepts were mentioned infrequently or in some cases absent all together as indicated by the letter ‘A’ in Table 2. While target concepts such as aortic valve regurgitation peak velocity, mitral valve regurgitation peak velocity, tricuspid valve mean gradient, and tricuspid valve orifice area were absent from the echo reports from WCM, Mayo Clinic and MIMIC reports, only aortic valve regurgitation peak velocity was absent from the Northwestern dataset. In addition, mitral valve mean gradient was absent from the WCM echo reports and mitral valve orifice area was absent from the MIMIC and Mayo echo reports. At WCM, aortic valve mean gradient, aortic valve orifice area, inter-ventricular septum dimension at end diastole, left atrium size at end systole, left ventricular dimension at end diastole and left ventricular dimension at end systole were also found in high frequency. Among Mayo, Northwestern and MIMIC echo reports, left atrium size at end systole was found frequently. Both WCM, Northwestern and MIMIC reports showed high recall and precision in extracting LVEF, with an F-score of 0.99, 0.99 and 0.95 respectively. We observed that EchoExtractor showed high precision (99%), while low recall (55%) for extracting LVEF from Mayo Clinic echo reports. Across all four datasets a F-score >0.90 was found on only two target concepts: aortic valve regurgitation and left atrium size at end systole. At WCM, Mayo Clinic and Northwestern, an F-score >0.90 was observed for aortic valve regurgitation, left atrium size at end systole, mitral valve regurgitation, and tricuspid valve regurgitation. Between WCM, Northwestern and MIMIC reports, a F-score >0.90 was observed for the concepts aortic valve regurgitation, aortic valve stenosis, left atrium size at end systole, left ventricular ejection fraction, and tricuspid valve regurgitation. Between WCM and Northwestern sites, mitral valve regurgitation showed an F-score >0.90.
<table>
<thead>
<tr>
<th>Concept</th>
<th>WCM</th>
<th>MAYO</th>
<th>NU</th>
<th>MIMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall (%)</td>
<td>Precision (%)</td>
<td>F-score</td>
<td>Recall (%)</td>
</tr>
<tr>
<td>aortic valve mean gradient</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>A</td>
</tr>
<tr>
<td>aortic valve orifice area</td>
<td>100</td>
<td>77</td>
<td>87</td>
<td>25</td>
</tr>
<tr>
<td>aortic valve regurgitation</td>
<td>96</td>
<td>100</td>
<td>98</td>
<td>87</td>
</tr>
<tr>
<td>aortic valve regurgitation</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>peak velocity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aortic valve stenosis</td>
<td>88</td>
<td>100</td>
<td>94</td>
<td>32</td>
</tr>
<tr>
<td>e/e prime ratio</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>inter-ventricular septum</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>38</td>
</tr>
<tr>
<td>dimension at end diastole</td>
<td>98</td>
<td>100</td>
<td>99</td>
<td>98</td>
</tr>
<tr>
<td>left atrium size at end systole</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>left ventricular dimension at</td>
<td>95</td>
<td>100</td>
<td>97</td>
<td>0</td>
</tr>
<tr>
<td>end diastole</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left ventricular size</td>
<td>67</td>
<td>97</td>
<td>79</td>
<td>73</td>
</tr>
<tr>
<td>left ventricular ejection</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>55</td>
</tr>
<tr>
<td>fraction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left ventricular posterior</td>
<td>6</td>
<td>100</td>
<td>12</td>
<td>67</td>
</tr>
<tr>
<td>wall thickness at end diastole</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mitral valve mean gradient</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>mitral valve orifice area</td>
<td>100</td>
<td>6</td>
<td>11</td>
<td>A</td>
</tr>
<tr>
<td>mitral valve regurgitation</td>
<td>96</td>
<td>100</td>
<td>98</td>
<td>91</td>
</tr>
<tr>
<td>peak velocity</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>mitral valve stenosis</td>
<td>60</td>
<td>100</td>
<td>75</td>
<td>17</td>
</tr>
<tr>
<td>pulmonary artery pressure</td>
<td>86</td>
<td>100</td>
<td>93</td>
<td>32</td>
</tr>
<tr>
<td>right atrial pressure</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tricuspid valve mean gradient</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>tricuspid valve orifice area</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>tricuspid valve regurgitation</td>
<td>90</td>
<td>100</td>
<td>95</td>
<td>92</td>
</tr>
<tr>
<td>peak velocity</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Error analysis of the echo reports from the study datasets revealed that false positive cases are mainly due to term mismapping or the reference range for a concept being wrongly mapped as the measured value. For instance, in WCM echo reports several target concepts are documented in semi-structured tabular format where a given concept is documented by the measured value followed by the reference value (e.g. “Aortic Valve Area 1.1 > 2.4 cm2”).
When there is no measured value documented as in several instances of aortic valve orifice area (e.g. “Aortic Valve Area > 2.40 cm2”) or left ventricular ejection fraction (e.g. “Ejection Fraction 0.55 - 0.75”) or mitral valve orifice area (e.g. “Mitral Valve Area > 3 cm2”), then the NLP system extracted the value given for the reference range as the measured value. For aortic valve orifice area and mitral valve orifice area, 23% and 94%, respectively, of reports were found to be false positive this way. Term mismapping also leads to a higher proportion of false positives. For example, the term “e:a” ratio being wrongly mapped to e/e prime ratio as in “Mitral E:A Rt 2” accounting 18% of false positive cases for the target concept e/e prime ratio. Although rare, the system wrongly mapped to terms “dilated aorta arch” to target concept left ventricular size with no apparent reason. In MIMIC echo reports, majority of false positive cases are due to term mismapping. Examples are “interatrial septum” being mapped to the concept interventricular septum dimension at end diastole, left ventricular cavity being mapped to the concept left ventricular dimension at end diastole, dilated descending aorta being mapped to left ventricular size, or mitral valve jet being mapped to tricuspid valve regurgitation peak velocity. Also, when a short representation of a target concept is part of a word preceded by a space character; the system wrongly mapped that concept. An example is “ava” for aortic valve orifice area being mapped to the word “available”. Also, some false positive cases were due to wrong mapping of the qualitative assessment of one concept to another. For example, the assessment “mildly thickened” of aortic valve leaflets was mapped to aortic stenosis as in “The aortic valve leaflets (3) are mildly thickened but aortic stenosis is not present.”

Error analysis on false negative cases revealed several instances where the system missed the target concept due to variations in the terms used to represent that concept in different clinical settings. For example, in WCM echo reports, the concept aortic valve max pressure gradient is consistently documented under peak aortic gradient, and since this is not one of the lookup terms for that concept and missed entirely. A second cause for false negative cases was when a line break (\n) or special characters (;, ~ , &), or parentheses appear between a concept and the corresponding value or qualitative assessment. Yet another cause for false negative cases was when a term for a target concept appears in two lines (e.g. right “\n” atrial pressure), and the system failed to extract the concept. Similarly, when two target concepts appeared too close to each other, the system failed to map those concepts with the correct value or assessment. For example, in MIMIC echocardiograms, there were several instances where aortic stenosis was described along with aortic valve orifice area (e.g. Mild AS (AoVA 1.2-1.9cm2)), and the system failed to correctly extract aortic valve area.

For the Mayo Clinic’s echo reports, we found that semi-structured table format had negative impact on the performance of the concept extraction. For example, “e/e prime ratio” listed in the table entries could not be detected consistently and accurately. Therefore, the semi-structured tables were excluded from the performance evaluation. Otherwise, the Mayo Clinic’s echo reports appeared to have relatively consistent syntax. We also found that semicolon (;) was used in place of colon (:) for the representation of some concept values such as “Calculated left ventricular ejection fraction; 65 %”, which resulted in roughly half of the false negatives being recorded for this concept. In addition, some phrase patterns appeared to be more commonly used at Mayo Clinic (e.g. Mitral valve sclerosis without stenosis), which resulted in 58% of the false negatives for the detection of the mitral valve stenosis concept and 47% of the false negatives for the aortic valve stenosis concept. A similar issue occurred for “Aortic valve systolic mean Doppler gradient 12 mmHg”, where the insertion of “Doppler” term caused a failure to detect the gradient concept. The high occurrence rate of these specific phrase patterns appeared to be institution-specific and needed to be handled in the future.

At Northwestern, we also found multiple semi-structured table formats in the echocardiogram reports, which also led to poor NLP performance for some of the target concepts. Also, we found many of the other sources of errors as in other datasets, including the interjection of spacing and special characters in the reports, omissions of spacing between concepts and values (e.g., Ao valve open2.2 cm), different concept terms used, and multiple concepts in the same sentence. Of the few false positives we found, the NLP did not recognize some less common abbreviations such as those for aortic insufficiency (AI) that were not listed in most reports. Many of the false negatives were due to the above reasons; in particular, multiple concepts were frequently missed as they were in a separate table and the description of the measures were partly abbreviated (vs. fully spelled out or completely abbreviated as in other reports). For example, left ventricular size measures such as left ventricular dimension at end systole were listed in the table as “LV Size-end systole” instead of “LVESD” or “left ventricular size-end systole.” The remaining few false negatives were due to missing values in sentences where other concepts were mentioned. For example, some reports started their summary section at the end of the report with summarizing multiple normal measures in 1 statement, such as “Left ventricular size, systolic function, wall thickness, and wall motion are all normal.” There were very few instances of
FNs where the value extracted by the NLP was the wrong value for the concept-value pair; thus, the NLP algorithm was good at finding the correct value when it found a concept.

Discussion

While many modern EHR systems record various concepts from echocardiograms in a more structured manner, semi-structured or unstructured echocardiogram reports are still the primary source in most clinical settings. In the past a number of NLP methods have been developed to extract various cardiac measures documented in echocardiography reports. While majority of these studies focused on the extraction of LVEF \[12,23-25\], some have attempted to extract a broader range of cardiac concepts \[21,26\]. However, to the best of our knowledge these methods have not been tested for their performance on data originated from clinical sites other than the original NLP algorithm and method development site. As a result, “true portability” of the developed NLP systems has never been rigorously assessed. The present study fill this knowledge gap by adopting a system developed initially by the VA, and then subsequently implemented at three non-VA medical centers without any modifications, to test its performance in processing echocardiogram reports.

In particular, the NLP system under investigation - EchoExtractor - was designed to extract 27 concepts from narrative text, echocardiograms, and radiology reports, and the original developers reported an average precision of concept-value extraction of 0.982 in echo documents, 0.969 in Radiology, and 0.936 in narrative text, with F-score of 0.844, 0.877, and 0.872 respectively. Specifically, the system reported precision and recall values for 20 data elements in echocardiograms. The remaining 7 target concepts were either absent or too low in total mentions to report performance concepts. Out of the 20 target concepts in echocardiograms, 17 elements have been reported with a precision of 0.90 or higher\[21\]. When compared to the reported data, the overall performance of the NLP system was found to be mixed. For an NLP system to be portable, a low variability in the recall and precision values across sites is needed. As shown in Table 2, this is true for some target concepts but not for others. For those concepts where we found low inter-column variability, EchoExtractor can be deemed as an effective NLP tool. In WCM dataset, target concepts such as inter-ventricular septum dimension at end diastole, left ventricular dimension at end diastole, and left ventricular dimension at end systole are documented as part of a tabular format within echo reports and appeared consistently across all reports, and hence resulted high performance for these concepts. On the other hand, concepts such as aortic valve orifice area and mitral valve orifice area, although appear in a tabular format and are typically documented along with a reference range in the echo reports. For these concepts we observed that the NLP system often failed to correctly identify the measured values from the reference range, thereby resulting in lower precision.

The Mayo Clinic dataset comprised of semi-structured tables to represent the echocardiography concepts. While the table formats had modest variations over the years, the NLP system was generally unable to detect concepts located in these tables successfully. Our evaluations suggest that the representation of target concepts using a table format is likely to decrease the NLP system performance, and alternative approaches to capturing data in these sections should be evaluated. Specifically, we observed 16 target concepts in the Mayo Clinic dataset, out of which the NLP system achieve a high precision and F-score >0.90 for only 4 concepts. A notable exception in Mayo Clinic dataset compared to other datasets is the low recall observed on LVEF. While WCM, Northwestern and MIMIC echo reports showed high recall (99%, 97%, and 91%, respectively), Mayo Clinic echo reports showed only 55 percent. On further analysis, we concluded that the high rate of failure to detect LVEF is due to a semicolon ‘;’ character (e.g. ejection fraction; 67) that appeared between the concept and value, which was causing system error. Common institutional phrasings such as “sclerosis without stenosis” that do not fit into the predefined patterns can be a significant source for concern and poor performance.

We observed 23 target measurements on Northwestern echo reports, out of which the NLP system showed a precision and F-score values >0.90 for 10 elements. Unlike echo reports from other sites, the concept tricuspid valve regurgitation peak velocity was found with high recall and precisions. Although this concept was mentioned infrequently in reports from other datasets, they were mentioned in ~50% reports at Northwestern using system recognized concept names. Similarly, the measurement tricuspid valve orifice area is found only in Northwestern echo reports, and extracted with a high recall and precision.

Unlike the echocardiograms from WCM, Northwestern and Mayo, MIMIC reports are entirely in a free-text format. Of the 16 target concepts found in MIMIC dataset, NLP system achieved a high precision and F-score values for 6 concepts. Majority of false positive and false negative cases in datasets that are more semi-structured arise out of differences in local document formats. Thus, the differences in performance between MIMIC and other datasets indicates a possible correlation between structure and tool usefulness. This is particularly important, since an ideal NLP system should be able to perform well when analyzing unstructured narratives. For those concepts observed with
low performance, it is also possible that physicians entering data in MIMIC are constrained by internal guidelines to use peculiar phrasing or punctuation and the NLP system missed to correctly identify those concepts.

Conclusion

The work described in this paper details a case study to adopt a specific instance (EchoExtractor) of a NLP system (Leo) previously developed by the VA. The adopters (WCM, Mayo Clinic and Northwestern) individually led the NLP system implementation without any modification, drawing on existing resources, and employing conventional software skills. The study demonstrates that concept-value extraction from echocardiograms can vary depending on local text formats and variations in lexical terms used to document various concept across different clinical settings. The NLP system used a custom lookup dictionary as part of its extraction algorithm\(^2\), and our analysis showed that the performance of the system for several of the target concept can be improved significantly by extending this lookup dictionary with additional terms specific to individual sites. Many of the false negative cases resulted because of the absence of terms in the custom dictionary. Despite the fact that this system was developed based on documents originated at thousands of points of care and authored by a multitude of clinical professionals within the VA, it also suffers from some of the same drawbacks as other NLP systems have reported including the need for a local customization of extraction logic and extending the lookup dictionary based on local data source.

Acknowledgements

This work was supported in part by NIGMS grant R01GM105688, NHGRI grants U01HG008673 and U01HG6379, NIH grant U01HG009450, and NCATS grant UL1TR001422. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


Using Priorities of Hospitalized Patients and Their Caregivers to Develop Personas

Elena Agapie¹, Logan Kendall², Sonali R. Mishra¹, Shefali Haldar¹, Maher Khelifi¹, Ari Pollack¹, Wanda Pratt¹,
¹University of Washington, Seattle, WA; ²Palantir Technologies, Washington, DC

Abstract
Hospitalized patients and their caregivers often access technologies like patient portals to understand what happens during their hospital stay. Although this access can lead to more patient engagement and positive health outcomes, many find that the technology does not support their needs. As a first step toward improving patient-facing technologies we create personas for hospitalized patients and their caregivers by following the Q Methodology, a technique for quantifying subjective opinion. We clustered 28 hospitalized patients’ and 19 caregivers’ attitudes towards receiving and managing information and working with their care team. We contribute three patient personas: patients who are (1) accommodating information seekers, (2) involved safety guardians, and (3) self-managing participators. We identify three caregiver personas: (1) cooperative information seekers, (2) vocal participators, and (3) hands-off safety guardians. These personas can inform future tools designed to support communication and information management for hospitalized patients and caregivers.

Introduction
Allowing patients to access their health information while in the hospital enables them to be active participants in their care, to have increased satisfaction, to engage in healthy behaviors, and to have improved understanding of their care. Patients want to understand many different aspects of their care, including their medications, the plan of care, and the workflow of activities. However, patients and caregivers (i.e., friends or family members who help or accompany the patient during their hospital stay) face substantial barriers to receiving information and engaging in their care while in the hospital. These challenges include the lack of supportive tools, poor setup of their hospital room, challenges communicating with their care team, cognitive challenges from treatment as well as their illness, and ergonomic limitations due to their environment.

Research thus far has highlighted the needs of patients and the difference between their needs and the needs of clinicians. Missing from existing work is an understanding of how patients’ needs differ from each other, and how patient-facing technologies should be designed to accommodate those differences. Researchers have found that hospitalized patients differ in their interest and approach to engagement with their care, and that people managing a chronic disease differ in their use of technology. Yet, more work is needed to understand how hospitalized patients’ and caregivers’ values and priorities differ, and how the design of patient-facing technologies can better support those differences. A common approach to designing for such different types of users is to use personas. To identify personas in order to support the information needs of hospitalized patients, we use the Q methodology and semi-structured interviews. We identify hospitalized patient and caregiver user groups based on how they prioritize information in the hospital and how they want to work with their care team. We base these user groups on our study with 28 hospitalized patients and 19 caregivers. Through this work, we contribute (1) an explanation of six distinct personas that represent hospitalized patients and caregivers, and (2) implications for the design of tools that support the varied information and communication needs of hospitalized patients and caregivers.

Related Work
In a hospital environment, patients and their caregivers find it challenging to maintain up-to-date awareness of the patient's health status and plans of care, managing their information and communicating with their care team. We briefly summarize this research and how creating personas for patients and caregivers contributes new insights.

Patient and Caregiver Information Needs in the Hospital
Patients want to receive a variety of information about their care such as the plan of care, the schedule of activities, information about their vitals, food, and providers on their care team. Patients also want to know that information such as their allergies or preferences is effectively communicated to their care team. They want to ensure that information they communicate is properly propagated to the care system, for example that their current clinical team
knows their allergy information\textsuperscript{15} or contextual information about their treatment routine\textsuperscript{16}, and that such information is properly transferred to clinical staff from one shift to another to achieve continuity of care\textsuperscript{16}.

Caregivers play an important role in managing care. Caregivers act as a proxy in the care of patients to make treatment decisions\textsuperscript{17,18}, or to communicate on behalf of the patient\textsuperscript{17}. They also provide social support in transition periods or when the patient faces cognitive challenges\textsuperscript{16}, normalize the patient experiences\textsuperscript{19,20}, and act as companion, assistant, representative, navigator, and planner for pediatric patients\textsuperscript{17}. In a critical care environment, patients and families have limited cognitive resources to be able to process and manage their health information. Patients may forget 40-80\% of what was said during a medical consultation almost immediately\textsuperscript{21}. The order and modality in which information is presented, perceived importance, and health context can all impact the patient’s likelihood of remembering information\textsuperscript{21}.

In addition to needing to receive a variety of types of information, patients and their caregivers want to effectively manage, capture, and communicate information to the care team\textsuperscript{16}. Without tools to accomplish this, patients would have to remember all the information providers share and be responsible for keeping their providers aware of their needs during shift changes and transitions in care\textsuperscript{16}. Further, critically ill patients often need substantial help from caregivers for managing information from clinicians\textsuperscript{22}.

This prior work has helped identify the multitude of information needs that patients and caregivers have. Our study adds to this knowledge by investigating how patients and caregivers prioritize different information needs and how they cluster together based on shared priorities.

\textit{Patient and Caregiver Communication with the Care Team}

The Institute of Medicine has emphasized that providers should engage in patient-centric care, “responsive to individual patient preferences, needs and values and ensuring that patient values guide all the clinical decisions”\textsuperscript{23}. Even though patient and caregiver engagement in the hospital is important, it can be heavily impacted by the relationship they have with their care team. A high-quality relationship with the provider might make people more comfortable communicating certain questions to providers, such as information seeking questions about medication\textsuperscript{24}, and more interested in engaging with their care team around improving safety issues\textsuperscript{9}. A good quality relationship is associated with lower disease burden and greater uses of preventative care services\textsuperscript{25}. When patients have an increase in perceived control and communication with the provider, it leads to improvements in trust in their provider\textsuperscript{26}. But when the relationship with the care team is not as strong, patients encounter significant barriers in communicating with their care team, such as asymmetrical power dynamics that are typically dominated by the physician\textsuperscript{27}, fear of being perceived as a “difficult” patient\textsuperscript{27}, and lack of comfort asking challenging questions such as asking if a health care provider washed their hands\textsuperscript{9}.

The patient’s level of engagement in their care and their attitudes to their care team play a role in the information they receive in the hospital and how they engage with it. In this work we contribute empirical findings about how patients and caregivers prioritize and balance relationships and communication with their care team relative to their information needs.

\textit{Personas in Health Domains}

Although some technologies exist to support patients’ needs for information about their care and engagement with the care team\textsuperscript{14}, existing tools insufficiently support tailored, patient-specific information\textsuperscript{14}. To understand how technologies might offer tailored support to patients and caregivers with different needs, we seek to identify different patient and caregiver personas based on their information and collaboration needs. Personas are user archetypes that characterize users based on their attitudes, goals and behaviors\textsuperscript{11}. They are informed by a variety of types of data and methods, from qualitative insights on user goals\textsuperscript{11} to data driven personas based on behaviors\textsuperscript{33}. Personas are used during the design process to help: build empathy towards users\textsuperscript{10}, facilitate communication about design choices\textsuperscript{10}; reduce conflicts about what the perceived user goals are\textsuperscript{10,11}, and summarize data about users and their needs. Personas have helped characterize users in many health domains such as heart failure self-management\textsuperscript{28}, home healthcare technologies\textsuperscript{29,30}, child cancer survivors\textsuperscript{31}, or older adults\textsuperscript{32}. Based on the demonstrated usefulness of personas in these health domains, we characterize users of patient facing technologies based on attitudes towards information and communication needs.

\textit{Methods}

We used a mixed methods approach to identify the attitudes and opinions of patients and caregivers around information needs, information engagement, information management, and the relationships and social perceptions surrounding care delivery in the hospital. We used the Q methodology\textsuperscript{14} to structure data collection with the patient and caregiver, and interviewed the participants twice, first during the study session, and second during a follow up a week after discharge. The study was approved by the researchers’ institutional review board.
**Q Methodology**

To understand how patients prioritize and make sense of getting and managing information while hospitalized, we used the Q Methodology\(^{12,34}\) to structure our data collection and analyze our data. The method quantifies subjective opinions through two key features—a set of statements about a topic, and a grid for ranking those statements in the shape of a normal distribution (Figure 1). By quantifying subjective opinions, the Q Methodology leads to the creation of personas by understanding what people feel most strongly about. It provides many of the advantages of qualitative methods, such as discovering themes in rich subjective data. The Q methodology is a small-sample technique that offers statistical validity to the qualitative interpretation of subjective data, making it useful for small samples (<50 people)\(^{12}\). The Q methodology has increasingly been used to cluster types of users based on beliefs and reactions to technology\(^{12}\), and to describe the tradeoffs and attitudes individuals have about health technology in a structured way\(^{9,13}\). This method showed how common attitudes towards health and technology supersede typical demographic segmentations like gender, age, and ethnicity in the context of a chronic disease population\(^8\).

The Q methodology requires participants to order a set of statements relative to one another using a grid like the one shown in Figure 1. Researchers created a set of 34 statements, as described below. To identify patient priorities, we use the Q sorting activity: a highly structured data collection activity in which participants need to organize the subjective statements in a grid shaped like a normal distribution. The extremes of the distribution are the statements that participants agree or disagree with the most, with Neutral in the middle. The statements placed at the extremes receive the most weight in the quantitative data analysis. Based on this distribution, we can compare each person’s ranking with the statistical technique of factor analysis, and thus, identify patterns of subjective opinions.

Participants were asked to read through the statements and place them into a grid based on how much they agree or disagree with the statement relative to all the other statements. The researchers then probed the participants to explain their reasoning for the placement of statements and their general opinion about the statements. After participants organized all the 34 statements into the grid, we interviewed them about topics such as their routines during their hospital stay, communication with staff, families and friends, or the use of technology while hospitalized. A follow-up phone interview was conducted within 7 days of the discharge. The patients and caregivers conducted the Q sorting activity and interviews separately, then discussed together, and took breaks as needed. The information from interviews is used to complement participant agreement with statements by adding qualitative insights.

**Creating the Q Statement Set**

For the Q sorting process we started with a set of 89 statements that described different attitudes related to interacting with information and people in the hospital environment, based on prior research on patient-provider communication, patient information needs, and based on data from prior surveys on patients and caregivers\(^4,14\). These statements pertained to communication preferences, empowerment, information management, information needs and relationships. Based on pilots with children and adults, we narrowed down the Q set to 34 statements [link to the Q set](https://patientsassafeguards.ischool.uw.edu/wp-content/uploads/2019/07/AMIA19Q.pdf), such as “(6) It’s important to speak up to doctors and nurses when I have a question”, “(3) It’s important to know why each medication is given to me”, or “(12) It’s important that I am not a bother to my doctors and nurses”. Caregiver statements were almost identical, with adjustments to the pronoun where appropriate. Since caregivers and patients can play different roles in care, using identical Q statements between these two populations allowed us to find where caregivers and patients converged and diverged from each other.

**Analysis**

Each participant’s ranking of statements was transcribed and entered into the freely available PQMethod software\(^{35}\). We separated the Q rankings created by patients from those created by caregivers. We used principal component analysis and rotated using the varimax method\(^6\). This analysis surfaces unique clusters where participants’ Q statements clustered together. Each identified cluster was therefore a selection of participants who unambiguously load highly in the cluster. Each cluster had a set of distinguishing statements that characterized a unique point of view, ranked from -4 (Disagree the most) to +4 (Agree the most). The distinguishing statements for each cluster are statistically significant, meaning that participants ranked those statements much higher or lower than participants in other clusters, at minimum threshold of p<0.05. In the results we will refer to the distinguishing statements by marking...
their relative ranking with an asterisk. For example, we will refer to knowing information about medication (+4*) to refer to a statement that participants strongly agreed with, is a statistically significant (at p<0.05), and is a distinguishing statement for that cluster. To supplement the quantitative analysis, we analyzed interviews qualitatively based on codes grounded in the original statement categories, with additional codes developed based on emergent themes identified during the content analysis.

**Participants**

We recruited 28 patients and 19 caregivers at two study sites: an adult tertiary care hospital and a pediatric hospital. Patients (13 male, 14 female) had a median age of 53 for adults, and 12 for children. Caregivers (2 male, 15 female) were of age: one was 18-29, seven were 30-39, six were 40-49, three were 50-59 years old, and two did not specify the age. Participants had diverse educational and racial backgrounds: Hispanic/Latino (6), American Indian/Alaska Native (2), Asian (5), Native Hawaiian/Pacific Islander (1), Black/African American (7), White/Caucasian (30).

Based on regional population demographics, we used purposeful sampling to identify eligible participants based on age, gender, race, ethnicity, and we sought an even distribution across hospital services (medical or surgical). Research coordinators approached patients who had been in the hospital for at least 24 hours, spoke English, and were able to participate in a conversation about their current care. The coordinators obtained informed consent, and conducted a verbal questionnaire to collect demographics and information about the current hospital stay. In our pediatric cohort, participants needed to be at least 7 years old and have parental consent as well as their own assent to participate. Caregivers included family members and friends who actively observed and participated in the care for the patient.

**Results**

The Q-sorting principal component analysis surfaced three clusters, from the PQMethod software analysis, for patient attitudes and three clusters for caregiver attitudes. We present these clusters described as personas in the next section. Of the 28 patient participants, two were not able to complete the Q sort process because of interruptions related to their health condition or hospital stay. Adult patient quotations are indicated by an “A” and youth quotations by a “Y”, followed by a number (e.g. A02), or “C” for caregiver, followed by the participant number (e.g. C08). A complete list of the distinguishing statements is available at https://patientsassafeguards.ischool.uw.edu/wp-content/uploads/2019/07/AMIA19Q.pdf.

The personas are different between the patient and caregiver groups, however patient and caregiver personas also share dimensions. Both patients and caregivers have safety guardian personas, but they differ in how engaged they are with the care team. Both patients and caregivers have information seeker personas, but they differ in how cooperative or accommodating they are with the care team. Both populations have personas that are interested in being active participants in their care, either holistically, or focused specifically on self-management.

**Patient Personas**

Patients fell into three clusters with shared attitudes towards managing and receiving information in the hospital, and collaborating with the care team. These patient clusters account for 45% of the variance in the patient group, and were relatively evenly split between Cluster 1 (n=9), 2 (n=6), and 3 (n=7). Four patients did not have significant associations with any of the clusters, or fit across multiple clusters, so they are not included in the persona descriptions below. The clusters had an even distribution between hospital sites (adult and children’s), gender, and type of service (medical vs. surgical). All patient participants cared about actively engaging with their care in that they strongly disagreed with statement #30: "I prefer to leave decisions about my medical care up to my doctor". The patients expressed general agreement with statements about wanting to know what will happen with their care for the day (#2), the importance of speaking up when they have a question (#6), and feeling like their doctors and nurses listen to them (#21). However, key differences emerged regarding how the participants prioritized types of information, relationships, and level of engagement. Based on these differences, we describe three patient personas.

**Patient Persona 1: The Accommodating Information Seeker**

The persona emerging from the first cluster of the Q analysis (n=9) is distinguished by 8 statements, summarized in Table 2. We call this persona the “Accommodating Information Seeker”, based on their priority for knowing different types of information as well as their apathy in learning more about or interacting with their care team.

The “accommodating information seeker” strongly values getting information about various aspects of their care, with an emphasis on information immediately relevant for them. For example, accommodating information seekers think it important to know about their medication (+4*), about rapidly changes of situation like the severity of their condition (+4*), and the way to get information when they are worried (+3*). Many patients spoke specifically about wanting to know changes in their medications. For example: “knowledge is power and [knowing] what medications I'm taking
These patients are “accommodating” to their care team in the sense that they are willing to let their clinicians go about their business without interruption. Accommodating information seekers found it less important to engage with their care team and know either the care team’s workflow (-3*), or their names and responsibilities (-2*). Despite patients’ interest in getting access to information when they are worried, patients did not think it important to feel like clinicians had time to answer their questions (-1*). Nonetheless, they were not concerned about being a bother to the care team (+1*) and were willing to assert their voices about topics they cared about, like medication (+4*): “If you’re still confused [and] nobody can help now, you worry. So you need something, like ask doctor what I have to do to keep my body continue to working” (A02).

Compared to other clusters, participants associated with this cluster felt that cost concerns and respecting privacy were least important. In contrast to the “accommodating information seekers”, patient participants in this group were less concerned about understanding the seriousness of their medical situation.

Patient Persona 3: The Self-Managing Participator

The third persona comprises participants in the third cluster of the Q analysis (n=7), and is distinguished by 9 statements, summarized in Table 2. We call this persona the “Self-Managing Participator”, based on patients’ interest in being active participants in their care and performing self-management of their condition.

Patients associated with this cluster placed greater emphasis on “active” participation (+4) as one of the highest ranked items in their sorting grid compared to the “Accommodating Information Seekers” (+1) and the “Involved Safety Guardians” (0). These patients found it important to know the reasoning of their doctors (+3*), and they were not concerned that they might be a bother to their care team (-4*). Patients in this cluster were also interested in the longer-term trajectory of their care, indicating “self-management” interest, such as when they would leave the hospital (+3*). One patient stated: “I’m a control freak, I like to be in control so it’s important that I am, if not totally in control, ... you really want to be an active participant so that you understand the consequences later on, because you expect to be living the consequences later on” (A05). These patients cared about aspects like illness severity (+4*), similarly to the “Accommodating Information Seekers”. But this cluster focused primarily on how their illness would impact the future: “I don’t want to know if it’s worsening, but I want to know how serious it is so I can do stuff myself to help.” (Y12).

Compared to the other clusters, patients in this cluster were least concerned about the way information is communicated (0*), possibly because they appreciate information in any form. These patients rated the question about

<table>
<thead>
<tr>
<th>Patient Personas</th>
<th>Distinguishing Statement</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodating Information Seekers</td>
<td>(3) “It’s important to know why each medication is given to me”</td>
<td>+4</td>
</tr>
<tr>
<td></td>
<td>(22) “It’s important to know how my doctors and nurses are communicating with each other at shift changes”</td>
<td>-3</td>
</tr>
<tr>
<td>Involved Safety Guardians</td>
<td>(25) “It’s important to know what kinds of safety errors can happen with patients in my situation”</td>
<td>+3</td>
</tr>
<tr>
<td></td>
<td>(11) “It’s important to get information in a way that I can understand”</td>
<td>+4</td>
</tr>
<tr>
<td>Self-Managing Participator</td>
<td>(17) “It’s important that I be an active participant in my healthcare”</td>
<td>+4</td>
</tr>
<tr>
<td></td>
<td>(12) “It’s important that I am not a bother to my doctors and nurses”</td>
<td>-4</td>
</tr>
</tbody>
</table>

Table 2. Example distinguishing statements for each attitude cluster, (P>0.05, +4=Most Agree, -4=Most Disagree)

and for what condition and is this a medication that’s just temporary related to this condition that I have or is this one of my regular medications that I take for a different condition? “ (A06).
Concerns with possible safety errors the lowest (-3), compared to the "accommodating information seekers" (+1) and the "involved safety guardians" (+3).

**Caregiver Personas**

Using the Q sort analysis, we identified three caregiver personas from the 19 caregiver participants. Our caregiver sample size was determined by the caregiver presence at the research site. All pediatric patients had at least one caregiver present (16 participants), but far fewer adult patients in this study had such active caregivers (3 participants). The three caregiver clusters identified in the Q analysis account for 59% of the variance in the caregiver group. Three participants fit across multiple clusters, so their data is not included in the analysis of the personas below. Each cluster included caregivers from the adult and the pediatric patients.

In contrast to the patients, the caregiver group had a greater consensus in attitudes, both agreement and disagreement, about the Q statement ranking. Thus, fewer statements distinguished each caregiver cluster. Some topics that caregivers commonly agreed upon included (1) giving importance to having information that's easy to share with others (#24), and (2) having someone else help keep track of things is important (#19). Caregivers also said they wanted to feel doctors had time to answer their questions (#14) and to know about different treatment options (#7). Caregivers did not agree with the statement that they should know only what the doctors think is important (#31).

**Caregiver Persona 1: The Cooperative Information Seeker**

The persona resulting from the first caregiver cluster of the Q analysis (n=10) is distinguished by 8 statements, summarized in Table 3. We call this persona the "Cooperative Information Seeker" based on the caregiver’s interest in specific details and information about the care being given to the patient and their value of getting along with members of the care team.

These caregivers were “information seekers”, wanting to know how serious the patient's condition is (+4*), having medications explained to them (+4), and knowing different treatment options (+3). These caregivers also found it important to be an active participant in the patient care (+2). These caregivers disagreed with “only wanting to know what the doctor thinks is important” (-4), or leaving the decisions about the patient to the care team (-2). These caregivers wanted to be “cooperative” with the care team, to have a positive relationships with the care team (+2*), and, that it is important to feel like doctors have time to answer their questions (+3). One caregiver placed the responsibility of a good relationship to the care team, instead of the patient-caregiver team. In a statement reflecting their own role as patient proxy, they said: “a bad start is a bad start. And that's why it's good rapport, and I think as a patient, I don't need to worry about that. Nurses and doctors need to” (C08).

This caregiver cluster is different from the other caregiver clusters in terms of interest in tracking. They expressed greater agreement with statements around the importance of information tracking (0), compared to the second cluster, the “Vocal Participators” (-3), and the third cluster, the “Entrusting Safety Guardians” (-3). These caregivers also strongly disagreed with relying on memory (-3), compared to the “Vocal Participators” (+1) and “Entrusting Safety Guardians” (0). The information seeker caregivers were relatively neutral in their attitudes about information management, compared to cluster 2 and cluster 3 who strongly disagreed with statements about tracking (#18).

**Caregiver Persona 2: The Vocal Participator**

The second caregiver persona is based on the second caregiver cluster of the Q analysis (n=3), and is distinguished by 8 statements, summarized in Table 3. We call this persona the “Vocal Participator” because the place the greatest importance on the idea of being an active participant.

<table>
<thead>
<tr>
<th>Caregiver Personas</th>
<th>Distinguishing Statement</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperative Information Seekers</td>
<td>(15) “It’s important to know how serious the patient's illness is”</td>
<td>+4</td>
</tr>
<tr>
<td></td>
<td>(27) “It’s important that I have a positive relationship with the doctors and nurses”</td>
<td>+2</td>
</tr>
<tr>
<td></td>
<td>(20) “I like to keep track of things by just paying attention and remembering”</td>
<td>-3</td>
</tr>
<tr>
<td>Vocal Participators</td>
<td>(17) “It’s important that I be an active participant in the patient's healthcare”</td>
<td>+4</td>
</tr>
<tr>
<td></td>
<td>(12) “It’s important that I am not a bother to the doctors and nurses”</td>
<td>-4</td>
</tr>
<tr>
<td>Hands-off Safety Guardians</td>
<td>(25) “It’s important to know what kinds of safety errors can happen with patients in our situation”</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(27) “It’s important to know how serious the patient's illness is”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(21) “It’s important that I am not a bother to the doctors and nurses”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(19) “I prefer to leave decisions about the patient’s medical care up to the doctor.”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(32) “It’s important to get information at a time when I can focus”</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Example distinguishing statements for each attitude cluster, (P>0.05, +4=Most Agree, -4=Most Disagree)
Caregivers in this cluster are “Vocal participators”. Across clusters, caregivers ranked statement #17, about being an active participant in the patient’s healthcare as the most important statement in the entire Q set (+4*), but participants in this caregiver cluster ranked it significantly higher than those in other clusters. They agreed with other statements such as being listened to by doctors and nurses (+4*), or speaking up when they have a question (+3*). This group strongly disagreed the most with the statement about leaving decision-making solely up to the clinicians (-4*). One caregiver summarized this sentiment: “I'm very active in her care, so I really don't prefer to leave any of it all the way up to the doctor. Her doctor, specific doctor who we don't get to see all the time because he's not on call, but always asks me my opinion first because he knows that I read up on the disease” (C03). These caregivers place importance on knowing the names and responsibilities of the doctors and nurses (+2*), and were not concerned about being a bother to their care team (-4*).

**Caregiver Persona 3: The Hands-off Safety Guardian**

The third persona resulting from the third caregiver cluster of the Q analysis (n=3) is distinguished by 6 statements, summarized in Table 3. One of the most representative statements for these “safety guardians” is to know about safety errors (+2*), to understand the doctor's reasoning when making a decision (+4), severity of patient illness (+4), medication information (+3). We call this persona the “Hands-off Safety Guardian” based on their focus on concerns about safety but neutral in controlling the patient’s care or even receiving information at their own convenience.

However, the caregivers in this cluster are more hands-off regarding their role relative to clinicians. They did not strongly agree or disagree with statements about leaving decisions about the patient care to the doctors (0*), and knowing only what the doctor feels is relevant (-1*). These caregivers are more balanced in leaving decisions to the care team (0*) than the “Cooperative Information Seekers” (-2) and the “Vocal Participators” (-4), who explicitly prioritized not leaving decisions to the care teams. Consistent with a more hands-off engagement, are the distinguishing statements that these caregivers do not find it important to get information at a time when they can focus (-4*), or that the care team needs to listen to them (-2*).

**Discussion**

This study demonstrates that using a mixed method approach for clustering patients and caregivers can identify different persona types for how hospitalized patients and their caregivers want to engage with their care and prioritize information. Although other work has described personas for people managing chronic conditions—such as heart failure, or cancer—this work adds insights into the type of personas we encounter in the hospital environment, for both patients and caregivers. Our analysis contributes knowledge that individual patients and caregivers place different relative values on three types of needs: information seeking about their care (e.g. updates about care, information about self-management), patient safety, and the active engagement in self-care and communication with the care team. Our personas demonstrate how, while individual patients and caregivers may care about all three types of needs, they prioritize each of these needs differently. This difference in relative values means that patient-facing technologies must include features to support a variety of approaches to receiving and engaging with information and communication in the hospital. We describe a variety of features based on prior work that patient facing technology might include to support the multitude of priorities highlighted by the personas.

**Designing for Information Seeking**

Two of the personas we identified prioritize access to information over a strong active participation in their care, so technology should be proactive in providing information, without people having to seek it out. Inpatient-facing systems need to offer a range of detailed information, including the status of the patient, often related to the immediate care in the hospital, plan of care, medication, changes in their condition, or information about severity of the patient’s condition. For medication information, systems should provide patients with information about what the medicine is, why they are taking the medicine, how to take and how much of the medicine to take, or what alternatives exist.

Because both the patient (Accommodating Information Seeker) and caregiver personas (Cooperative Information Seeker) reflect a need to have a good relationship with providers or to not be a burden, information should be delivered in a way that is very easy for them to understand it. For example, the provided information could adapt to the level of health literacy of the patient and use delivery modalities like audio and video.

To prioritize caregiver’s needs, inpatient portals could support sharing features or the ability to have several caregivers interact with the portal. For example, technology should enable collaboration between several caregivers in tracking data about what is happening in the hospital, such as symptoms, or what the providers communicate about patient care. These features should be designed with special considerations for maintaining privacy between the patient and caregiver when appropriate.
The different priorities in being an Accommodating Information Seeker patient versus a Cooperative Information Seeker caregiver leads to different design implications for how these personas want to engage with the care team. Because caregivers are not as worried about being a burden to providers, systems can facilitate cooperation with the care team by sharing the names, photos, roles, and schedules of individual care team members. This can make it easier for the caregiver to engage with the team. To support the Accommodating patient, systems can enable the patient to communicate their needs while providing flexibility in the modality of this communication. Systems can prioritize asynchronous electronic communication, and flag which questions are urgent to the care team, to reduce patient concerns about being a bother and allow them to perform question “triage”\(^{40}\). To make patients more comfortable in contacting specific providers to ask questions, systems can include information about how each member of the care team prefers to be messaged, or the ability to choose which member of the care team they contact.

**Designing for Safety**

Being a safety guardian is often associated with being a vigilant monitor, in particular of communication with clinicians\(^{41,42}\). Systems should include a variety of features to support the different levels of involvement that Safety Guardians want to have. The needs reflected by the Involved Safety Guardian suggest that patient-facing portals should include features that promote active participation from patients, and enable keeping patients aware of safety issues, medication, and errors that can occur. Systems could support patients in actively reviewing their medication and preventing errors\(^{40,43}\). Through active engagement, patients could use checklists where to monitor whether they are getting the correct medication, alerts and reminders to monitor provider’s hygiene, and other potential hazards.

For the Hands-Off Safety Guardian caregiver persona, who is less engaged in managing safety issues, systems can provide ways to monitor their safety that do not introduce greater cognitive or physical burdens, such as FAQs about safety, common questions by patients like me, notification about important safety issues (e.g., changes in medication), or issues to watch for when taking medication. By including a variety of features, patient-facing technology can appeal to patients and caregivers with varied levels of engagement in care.

**Designing for Active Participation**

Being an active participator in one’s care involves a high interest in both information about care and in engaging with the care team, thus systems should be designed to actively support patients in participating in their care. To support the Self-Managing Participator patient, patient-facing portals should be proactive in offering information about the patient’s care and self-management of their condition, such as changes occurring in their care, changes in medication, severity of their condition, or plan of care. To help patients prepare for discharge and managing care outside of hospital, technology can help patients build up knowledge about how to manage their health, and techniques to build self-efficacy towards managing their health\(^{44}\).

The Vocal Participator caregiver persona prioritized a high level of engagement with the care team. Systems can support them with features that enable high engagement: requesting a second opinion, engaging with doctors and nurses at times that are convenient not only for the clinicians but also for the patient and caregiver, notifications about when their doctor is due to arrive for consultation, what each care provider is in charge of for their care, or the ability to give feedback to the provider about what they did well or what could be improved. This information is of importance to caregivers, because that can help them be present for important communication events such as during rounds or when important decisions are discussed.

**Limitations and Future Work**

Although personas are a commonly used tool to help designers make decisions, they will not capture every user’s needs; rather, personas are meant to identify archetypes of a variety of users. People might have additional needs beyond what has been identified in prior work\(^{41-44}\). In addition, our sample was designed for breadth rather than depth in understanding how the needs and values of patients and caregivers differ between hospital settings or units within a hospital. Our approach offers a broad holistic view of hospitalized patient and caregiver priorities, but it is possible that if medical and surgical patients’ data were analyzed separately there might be a different prioritization of needs across personas. Also, we assessed how patients and caregivers prioritized information and communication needs at one point in time, but patient or caregiver priorities could change during their stay in the hospital. Further research is needed to understand whether these personas can be applied broadly across all hospital contexts, or if certain care contexts bring with them certain priorities. We have found distinctions between the patient and caregiver personas, but more work is needed to understand how the differences in priorities of participants in our study transfer to different population groups, including based on demographics, disease type, settings, or across different types of caregivers.
Conclusion
Our work contributes a new understanding of information management and communication priorities of hospitalized patients and caregivers. Through our mixed-methods approach, we identified three patient personas: accommodating information seekers, involved safety guardians, and self-managing participators. We also identified three caregiver personas: cooperative information seekers, vocal participators, hands-off safety guardians. These personas could be used by designers of technologies to ensure that during the design process the personas facilitate design choices that support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs. We hope that versatile technology will support the needs of different kinds of people, with different priorities and needs.
35. Schmolck P. PQMethod Software. 2014.
Smartphone Monitoring of Mood Instability in Young Depressed Patients: A Latent-class Analyses

Vibha Anand, PhD¹, Bo Hu, PhD², Amit Anand, MD³

¹Center for Computational Health, IBM T.J. Watson Research Center, Cambridge, MA, USA
²Quantitative Health Sciences, Cleveland Clinic, Ohio, USA
³Center for Behavioral Health, Cleveland Clinic Foundation, Ohio, USA

Abstract

This study captured daily and weekly mood ratings using a smartphone from bipolar disorder (BD) and unipolar major depression disorder (MDD) subjects at high (HRMDD) and low risk (LRMDD) for developing Bipolar Disorder (BD) and healthy controls (HC). Method: 40 subjects (18 – 30 yr) (6 BD, 13 HRMDD, 16 LRMDD and 5 HC) were studied and a total of 2401 daily and 744 weekly ratings were collected. HRMDD and LRMDD subjects were naturalistically treated with antidepressants. We investigate if latent-class analyses of ratings can detect mood instability among MDD and BD groups. Results: Our analyses revealed four underlying mood states correlating with clinical mood states. There was a trend for greater number of state changes in BD and HRMDD subjects compared to LRMDD and HC groups. Conclusion: Smartphone ratings may adequately capture mood instability in BD subjects and at risk HRMDD subjects and offers a prudent way for monitoring development of serious manic symptoms.

Introduction

Depression can manifest itself in two forms – unipolar or major depression (MDD) in which subjects only suffer from episodes of depression (low mood and lack of pleasure, decreased energy, sleep/appetite problems, decreased concentration, hopelessness and suicidal ideation) and bipolar disorder (BD) in which patients suffer from periods of depression but also suffer from periods of mania (elation or irritability, excessive energy, lack of need for sleep, rapid speech, grandiosity, poor judgment and impulsivity). [1] BD is a more serious illness as it is frequently associated with more severe depressive symptoms and with mania that can lead to serious social and legal consequences. Furthermore, the mood instability from periods of depressive symptoms to periods mania can lead to an unstable and chaotic lifestyle.

One critical problem in the treatment of MDD, particularly in young patients, is that some MDD patients may have a hidden risk for developing a (hypo) manic episode and conversion of diagnosis to BD. Many of these vulnerable subjects have a family history of BD or have sub-threshold BD symptoms. [2,3] It is critical to closely follow-up these patients for changes in mood symptoms particularly during antidepressant treatment as the response to treatment can be unpredictable in these patients. Some studies suggest a propensity for precipitation of mania or rapid antidepressant response [4] while others suggest that antidepressants can be safely given to MDD groups at high risk of developing BD. [5] Clinical appointments spaced out over several months may not be able to capture fluctuations in mood or occurrence of (hypo)manic symptoms. Patient reports at these appointments with clinicians are at best subjective retrospective recollections compounded by lack of recognition of (hypo)mania symptoms by the patients in themselves.

Therefore, it is critically important to develop new methods to measure any increase in mood fluctuations in MDD patients that have been recently started on antidepressants. These new methods need to be more sensitive to changes occurring over time as well as be more objective in terms of how they measure changes in mood instability between periods of depressive, (hypo)manic and euthymic states. In this regard, the recent availability of smartphone technology and the emerging field of behavioral informatics has provided a new way of acquiring and measuring mood data. Smartphone applications are being increasingly used to monitor mood symptoms and also deliver
interventions in MDD and BD. [6–10] However, few studies have investigated at-risk mood disorder populations to monitor effect of antidepressant treatment using smartphone technology.

In this study, we used a smartphone application to monitor the effect of antidepressant on mood instability in terms of fluctuations between mania and depression, in young MDD subjects who are at a high risk of developing BD (HRMDD) and compared them to young patients with depression at low risk of developing bipolar disorder (LRMDD) as well as with subjects already diagnosed with BD and Healthy Controls (HC). We perform latent class analyses of smartphone survey ratings. The ratings represent time-series of survey scores collected from subjects while under treatment or from HC. Using a data-driven approach, we aimed to find latent states that may correlate with clinical mood states in these groups. Furthermore, we aimed to find group-level differences in their (latent) state transitions. Our hypothesis was that HRMDD group will be similar to the BD group in that they will show an increase in mood instability over time and will have higher spikes of (hypo) manic symptoms compared to LRMDD and HC groups.

**Methods**

Medication-free subjects were recruited from an ongoing study for treatment of MDD in young adults in which after a baseline assessment, open-label treatment with antidepressants was given and patients were followed up with periodic assessments and ratings of depression and mania scores. In addition to these procedures, the subjects were consented separately to take part in a smartphone mood monitoring study in which they could record their mood on a daily as well as weekly basis using a proprietary smartphone app developed by the company - Ginger.io. (San Francisco, California)

**Inclusion criteria for MDD subjects:** Ages 15-30 years and able to give voluntary informed consent; 2) Satisfy criteria for DSM-IV-TR Major Depressive Episode using a Structured Interview; 3) Never met criteria for mania or hypomania; 4) 17-item Hamilton Depression Rating Scale score (HDRS) [11] > 15 and < 25; 5) Young Mania Rating Scale (YMRS) [12] score < 10; 6) Able to be managed as outpatients during the study as ascertained by the following – i. Clinical Global Severity Scale [13] < 5 i.e. moderately ill, ii. No significant suicidal or homicidal ideation or grossly disabled; 7) Have a smartphone on which the ginger.io mood application can be used; 8) willing to record their mood and activity using the Ginger.io application. **Inclusion criteria for BD subjects:** 1) Satisfy criteria for DSM-IV-TR Bipolar Disorder Depression using a Structured Interview; 2) Rest of the criteria same as that for MDD subjects; 3) on no psychotropic medication. **Inclusion criteria for Healthy Control (HC) subjects:** 1) no personal or family history in first degree relative of psychiatric illness; 2) no significant neurological disorder; 3) no history of alcohol of substance dependence in the past 12 months; 3) on no psychotropic medication. **Ascertainment of HRMDD vs LRMDD is described in a previous publication.** [14]

**Exclusion criteria for all:** 1) meeting DSM-IV criteria for schizophrenia, schizoaffective disorder, or an anxiety disorder as a primary diagnosis; 2) use of psychotropics in the past 2 weeks; use of fluoxetine in the past 5 weeks; 3) acutely suicidal or homicidal or requiring inpatient treatment; 4) meeting DSM-IV criteria for substance dependence within the past year, except caffeine or nicotine; 5) positive urinary toxicology screening at baseline; 6) use of alcohol in the past 1 week; serious medical or neurological illness; 7) current pregnancy or breast feeding.

**Smartphone application:** The Ginger.io Behavior Platform as implemented at the time of the study was used by researcher groups, clinicians, patients and providers alike to collect data. In this paper, we are reporting results from active data collection using surveys and other self-reported data gathered through user input. This was achieved in two ways: (i) the user receives a notification that a survey is available and upon clicking the notification, they are transferred to the survey page, (ii) the user can then manually launch the application and they are shown all the surveys available at that time. The data described above is encrypted and transmitted over a secure 128-bit SSL 3.0 connection using the HTTPS protocol for transmissions between mobile application, web application and secure servers.

**Smartphone surveys:** Daily ratings were for the following items on a 1 – 5 visual analog scale with 1 signifying “Not at all” and 5 “Signifying Most of the day” on the other. The daily ratings for (hypomania) ratings were questions for the last 24 hours for the following items - 1) feeling much more excited or full of energy than usual; 2) feeling so good or so hyper that other people thought you were not your normal self or you were so hyper that you got into trouble; 3) feeling so irritable or grouchy that you either started arguments, shouted at people or hit people; 4) sleeping far less than usual and still not feel tired or sleepy. Weekly ratings were also done to assess bipolar symptoms for the past 1 week on the following items and a visual analog scale from 1 – 7 with 1 signifying “Not at all” on one end and 7 signifying “all the time” on the other. These surveys were done on the following items – 1) did you feel happier or more cheerful than usual; 2) did you feel more self-confident than usual; 3) did you sleep far less than usual and still not feel tired or happy; 4) did you talk more than usual; 5) were you more active either socially, sexually, at work,
**home or school than usual.** The mania items were summed up to get one score for daily and weekly ratings each respectively.

**Treatment:** after the baseline ratings MDD patients were started on antidepressant treatment usually with selective serotonin reuptake inhibitor antidepressant, most frequently fluoxetine, unless patient wanted another antidepressant. Patients were given “real world” open label treatment and dosage was adjusted depending on tolerance or response. Augmentation with another antidepressant such as bupropion or change to another antidepressant such as a dual uptake inhibitor e.g. duloxetine was also done depending on tolerance and treatment response. Bipolar subjects were followed up in the regular outpatient clinic and treated with antidepressants and mood stabilizers as indicated.

**Statistical Methods and Analyses**

Two types of analyses were done for measurement of mood instability: 1) *A priori* defined criteria; and 2) *Latent-class* analyses of weekly rating score time series using a data-driven approach.

**A priori criteria for mood instability:** The aim for these analyses was to measure daily and weekly sub-threshold (hypo)mania symptoms scores in a clinical setting – 1) a priori defined criteria for spikes in scores (>2 threshold and >25% increase from baseline) and fluctuations (change in two consecutive ratings of more than 25%). We also measured differences in subject groups using statistical methods (ANOVA).

**Latent-class Analyses:** using Hidden Markov Model (HMM) [15], mainly to test the hypotheses that the number of latent states in the *weekly mood ratings* correspond to the number of mood states inferred (using overall clinical ratings and assessment) by the clinician. We also hypothesized that the transitions of a subject in underlying latent states (as measured by percent times a subject switches states over the course of the study) will vary in expectation with clinical classification of mood states. Accordingly, HRMDD subjects would have the highest switch rate (as the BD subjects in the study were likely being treated with mood stabilizers).

**Modeling:** The aims of modeling are to a) discover the number of latent states that may represent clinical mood status in weekly mood ratings dataset; and b) gain insights as to the characteristics of these states in relationship to the smartphone surveys and subjects’ sub-groups or class label of HRMDD, LRMD, BD, HC. HC were studied to study the differences from normal for each of the patient group. In this regard, please note that the modeling did not target any particular variable from the smartphone survey or subject-level data.

For the first aim, we constructed several HMMs using the python library (https://pypi.org/project/hmms) and the weekly mood ratings data. The 5 distinct weekly rating questions (measured on a scale of 1 to 7) were summed to get a total score of 1 to 35 per subject for each weekly assessment. Then for learning a model and evaluating its performance, the dataset was randomly split into training and test (held-out) sets of weekly mood rating sequences (70-30% split). To address the need for learning a model from a larger sample or all available data, we use 10-fold cross validation approach, i.e. we used 10 random splits (samples of training and test sets) and performed independent experimental setups (as detailed below). This is in accordance with accepted machine learning practice of evaluating a model to avoid overfitting. [16]

Our primary interest was to find an explainable set of latent states that may correspond with clinical mood status (i.e. euthymia, depression, and range of mania symptoms). Thus, we limit our exploration to the 5 latent states (varied from 2 to 6). It’s possible that a greater number of latent states may increase the model fit but may also reduce explainability. Thus, we sought to compare 5 separate latent state models in the 10 experimental setups. We selected the number of latent states based on a model that maximizes the likelihood of the data (i.e. held-out set). This model can then be explored further for latent state characteristics and decoding of weekly mood ratings. Below we give details of the experimental setup, performance metric and model selection and model-based decoding of the weekly mood ratings data.

**Experimental setup:** In each experiment, we parameterize a model to learn a fixed (N) number of latent states (using the training set), among other model parameters (i.e. initial state, emission probabilities and transition matrix) for an HMM using the Expectation-Maximization (Baum-Welch) algorithm [17] implemented in the python library. To measure model performance, we compute the log-likelihood (probability) of data (i.e. total score of weekly mood rating sequence) for every subject in the held-out set under the model (parameters). We also start each experiment from random initialization (seed) to mitigate the issue of local maxima and repeat the above procedures for every latent state we wished to explore, i.e. we varied N from 2 to 6 latent states in all the 10 experimental setups. Overall,
we performed 50 experiments and computed log-likelihood of held-out set in each.

Performance metric and Model selection: Individual held-out subject probabilities were computed in each experiment, and log summed up for obtaining the log-likelihood (LL) of the (held-out) test set, i.e. predictive LL of the model. This score was used for model comparison of all the latent-states explored. In general, the higher the predictive LL score, the better the model fit to the number of latent states and the observational data (i.e. weekly mood rating sequences). In our setup, we selected a model based on the median value of the predictive LL for all the latent state explored in the 10 random experimental setups. The model that best fit the observational data (D) with the number of latent states (N) is the one with maximum median LL, i.e. argmax(P(D, N | Θ)), where Θ are the learned model parameters. Here we refer to this model as the selected model.

Model based decoding: The selected model was then used to decode an individual subject’s weekly mood rating sequence, i.e. the selected model assigned a latent state to every subject’s weekly mood rating. Here we use the state assignment to draw observational insights and its association to other subject-level characteristics, such as their (known) class label or sub-group. Please note that these analyses are post-hoc and sub-group label was not used for modeling in anyway. In particular, we examine percent (latent) state change by these sub-groups, i.e. number of times the latent state change occurs over the observational period which is defined as the number of subject-level measures. We also examine clinician interpretation of the modeling states and any age or gender differences in these states.

All data were collected in accordance with IRB approval. Cleveland Clinic Foundation IRB approved this study.

Results

A total of 2401 records from 40 subjects (6 BD, 13 HR or HRMDD, 16 LR or LRMDD and 5 HC) were available for daily mood ratings. A total of 744 records were available from 38 patients (6 BD, 13 HRMDD, 14 LRMDD and 5 HC) for the weekly ratings for 2 to 44 weeks. Please see Table 1 for subject characteristics.

Table 1: Study Subject Characteristics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>BD (N=6)</th>
<th>HR (N=13)</th>
<th>LR (N=16)</th>
<th>HC (N=5)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (mean (sd))</td>
<td>24.3(4.1)</td>
<td>22.9(3.6)</td>
<td>25.6(3.3)</td>
<td>23.6 (4.6)</td>
<td>ns</td>
</tr>
<tr>
<td>Gender (Female (n) %)</td>
<td>4(67)</td>
<td>12(92)</td>
<td>9(56)</td>
<td>1(20)</td>
<td>0.026</td>
</tr>
<tr>
<td>Race (Caucasian (n) %)</td>
<td>4(67)</td>
<td>7(54)</td>
<td>13(81)</td>
<td>3(60)</td>
<td>0.005</td>
</tr>
<tr>
<td>Years of illness (mean (sd))</td>
<td>11(5.4)</td>
<td>9.2(6)</td>
<td>10.9(5.2)</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Age at first episode (mean (sd))</td>
<td>14(3.6)</td>
<td>13.7(4.5)</td>
<td>14.7(4.5)</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Medication free period in weeks (mean (sd))</td>
<td>76.3(84.5)</td>
<td>66.7(73.4)</td>
<td>88.9(137)</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Number of Depressive Episodes (mean (sd))</td>
<td>14.2(16.4)</td>
<td>11.7(12.3)</td>
<td>41.1(45.6)</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Duration of medication in days (mean (sd))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ns</td>
</tr>
</tbody>
</table>

A priori defined percent mood fluctuations: For percent mood fluctuations between two consecutive ratings, daily ratings showed differences between groups (BD: 13%, HRMDD, 5%, LRMDD: 6%, HC: 0%) and this difference was significantly different only between BD and HRMDD (p = 0.02). For percent mood fluctuations in weekly ratings, (Figure 1), differences were seen between groups (BD: 25%, HRMDD: 15%, LRMDD: 10% and HC: 0%) and for patient groups this difference was significant between BD and LRMDD (p = 0.05).

A priori defined percent mood spikes: For percent of manic spikes, significant differences were seen between groups for daily ratings (BD: 35%, HRMDD: 6%, LRMDD: 11%, HC: 0%). This difference was significant between BD and HRMDD (p=0.001) and BD and LRMDD (p=0.05) but not different between HRMDD and LRMDD. For weekly ratings, differences were also seen between different patient groups (BD: 49%, HRMDD: 19%, LRMDD: 27%, HC: 2%). For daily ratings this difference was significant only between BD and HRMDD (p = 0.001).
Latent-class analyses of weekly mood ratings: The *predictive LL* was largest for the model with 4 latent states, followed by 5, 2, 3 and 6 latent state models (median values: -3.77, -4.28, -4.46, -4.49, -4.80 respectively). Thus, using a data-driven approach, a 4-state model was selected for further interrogation. A 4-state model also lends itself well to the clinician’s interpretation of mood status as euthymic, depressed, and range of mania symptoms. Therefore, further results here are derived from the best model among the 4-state models (in terms of predictive LL). Below we describe latent state changes as observed in the data.

**Percent latent states change**: Please refer to Figure 2 for the following text. As expected, healthy controls showed no changes (as reflected by latent states in our model), and LRMDD and HRMDD subjects show lower number of (median) state changes when compared to BD subjects.

![Figure 1: A priori criteria percent mood fluctuations](image1)

![Figure 2: Percent latent states change](image2)

*Note: BP refers to BD sub-group (Left), LR is LRMDD sub-group, HR is HRMDD sub-group (Right)*

We further observed the following latent state characteristics as in Table 2. It is likely State 0 corresponds to euthymic mood status as in a HC (i.e. negligible mania symptoms), State 1 corresponds to minimal mania symptoms status, State 2 is few mania symptoms mood status and State 3 is mild mania symptoms mood status. This classification can be made on the basis of higher mania symptom scores as seen in states 2 and 3 (Table 2).

**Table 2: Latent State characteristics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>State 0</th>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(mean, sd)</td>
<td>(±4.70)</td>
<td>(±2.11)</td>
<td>(±2.84)</td>
<td>(±3.92)</td>
</tr>
<tr>
<td>Total weekly score (1-35)</td>
<td>1.20</td>
<td>2.10</td>
<td>5.31</td>
<td>13.33</td>
</tr>
<tr>
<td>Feel happier</td>
<td>0.29</td>
<td>0.47</td>
<td>1.40</td>
<td>2.75</td>
</tr>
<tr>
<td>(±1.14)</td>
<td>(±0.75)</td>
<td>(±0.99)</td>
<td>(±1.27)</td>
<td></td>
</tr>
<tr>
<td>Feel self-confident</td>
<td>0.28</td>
<td>0.41</td>
<td>1.27</td>
<td>2.84</td>
</tr>
<tr>
<td>(±1.11)</td>
<td>(±0.70)</td>
<td>(±1.05)</td>
<td>(±1.24)</td>
<td></td>
</tr>
<tr>
<td>Sleep less than usual</td>
<td>0.16</td>
<td>0.40</td>
<td>0.67</td>
<td>2.13</td>
</tr>
<tr>
<td>(±0.76)</td>
<td>(±0.79)</td>
<td>(±1.18)</td>
<td>(±1.46)</td>
<td></td>
</tr>
<tr>
<td>Talk more than usual</td>
<td>0.21</td>
<td>0.31</td>
<td>0.77</td>
<td>2.74</td>
</tr>
<tr>
<td>(±0.98)</td>
<td>(±0.63)</td>
<td>(±0.96)</td>
<td>(±1.26)</td>
<td></td>
</tr>
<tr>
<td>Active more than usual</td>
<td>0.26</td>
<td>0.40</td>
<td>1.22</td>
<td>2.88</td>
</tr>
<tr>
<td>(±1.08)</td>
<td>(±0.69)</td>
<td>(±1.15)</td>
<td>(±1.35)</td>
<td></td>
</tr>
</tbody>
</table>
Furthermore, it is clinically known that subjects in the BD sub-group continue to have mild spikes even when on medications. [18] This is also depicted as higher latent state changes in the BD sub-group in Figure 2. We depict an example decoding of one BD subject’s weekly mood rating sequence (into latent states) in Figure 3 and table below.

A typical BD sub-group subject switches states over the course of time (3 times or 9% for the subject in Figure 3 – marked with * in Table below). We found that overall the mood status (as measured by latent states change) fluctuated up to a third (i.e. 33% of times) in all subjects studied. There was no state fluctuation in HC subjects, and on average 5% in LRMDD, 7% in HRMDD and 11% in BD subject sub-groups. This trend in latent state fluctuations was similar to that calculated directly using our ‘a priori criteria’ for mood score fluctuations, as mentioned above.

**Discussion**

The results of this study indicate the feasibility of assessing daily and weekly mood instability in subjects, particularly undergoing antidepressant treatment, using a smartphone. A few other studies have also reported good feasibility of capturing mania symptoms using daily ratings. [8] However, in this study, we also conducted weekly mood ratings. In general, daily mood and weekly mood ratings changes as measured with a priori criteria were in similar direction indicating that weekly mood ratings may suffice. Weekly ratings may decrease the burden on the subject in terms of answering daily surveys. [6]

In terms of differences among patient groups based on daily and weekly changes in mania symptoms, the BD group showed the greatest number of changes, the LRMDD group the least and the HRMDD group showed intermediate fluctuations (Figure 1) based on a priori criteria for spikes and fluctuations. Adding to face validity of these measures – the healthy controls showed little or no mania spikes or fluctuations. However, the difference between HRMDD and LRMDD for mood spikes and fluctuations, as defined, was not significant. This could indicate no difference between the two groups in terms of mood instability or that the a priori measures as defined were not constructed properly to show the difference. In terms of continuous measures, it is difficult to define mood spikes or fluctuations beyond consecutive ratings as many such measures could be constructed using 3, 4 or more number of ratings grouped at a time.
In this regard, conceptualizing mood and mood disorders as complex and non-linear systems can add to our current understanding of mood regulation [18] and a data-driven latent class model which explores hidden states can be much more objective and informative. The classification of mood ratings by this model also has face-validity as assessed by the clinician author (AA). Using a 4-state model, we found that both HRMDD and BD subjects have a greater number of state changes than LRMDD subjects. The direction of differences between groups (BD>HRMDD>LRMDD>HC) was the same using the ‘a priori criteria’ of fluctuations in mood symptoms and that for the latent mood states. These results suggest that latent class analyses of time series of mood states available from smartphone ratings provides a better analytical framework to identify mood instability.

Limitations of the study were the small number of subjects that were studied though a large number of data points were obtained. Future studies are needed with larger number of subjects particularly within each diagnostic subgroup. Comparison groups also need to be closely matched in terms of age, gender and ethnicity. Another limitation was that subjects started their smartphone ratings at different durations of treatment therefore future prospective studies are needed with all subjects at the same point of their treatment preferably at the start of the study. This study involved open label treatment with a variety of medications along with a naturalistic follow-up. Randomized controlled trials, possibly with placebo will be needed to further confirm the bipolar like mood instability in HRMDD patients while on antidepressant treatment. Smartphone ratings also need to be correlated with clinical ratings that are the gold standard in terms of mood ratings. Correlational analysis between smartphone ratings and clinical ratings along with clinical assessment of severity of illness need to be done to further validate the use of mobile ratings.

**Conclusion**

In summary, the results of the study indicate feasibility of doing smartphone ratings at a daily and weekly basis in mood disorder subjects. Weekly ratings may be adequate to decrease the rating burden on the subjects. In terms of clinical findings, bipolar disorder patients seem to have the most mood instability compared to MDD subjects. Within the MDD group, while on antidepressant treatment, subjects thought to be at high risk of developing BD (HRMDD) exhibited higher mood instability than those at low-risk (LRMDD). As the HRMDD subjects are at a risk of developing full-blown manic symptoms, weekly smartphone ratings may be used to monitor their mood to predict such events.

**Acknowledgements**

Part of this study was supported by NIMH grant to AA (R01MH093420).

**Address for correspondence**

Amit Anand, MD  
Center for Behavioral Health, Cleveland Clinic  
9500 Euclid Avenue P-57 Mailbox  
Cleveland, OH 44122  
Email: ananda@ccf.org  
Fax: 216-636-2995  
Phone: 216-636-2840
References


Towards Reliable ARDS Clinical Decision Support: ARDS Patient Analytics with Free-text and Structured EMR Data

Emilia Apostolova, PhD1, Amit Uppal, MD2, Jessica E. Galarraga, MD, MPH3,4, Ioannis Koutroulis, MD, PhD5, Tim Tschampel6, Tony Wang PhD7, Tom Velez, PhD, JD6
1Language.ai, Chicago, IL; 2NYU School of Medicine, Bellevue Hospital Center, New York, NY; 3MedStar Health Research Institute, Hyattsville, MD; 4MedStar Washington Hospital Center, Georgetown University School of Medicine, Washington, DC; 5Children’s National Health System, Washington, DC; 6Computer Technology Associates, Ridgecrest, CA; 7Imedacs, Ann Arbor, MI

Abstract

In this work, we utilize a combination of free-text and structured data to build Acute Respiratory Distress Syndrome (ARDS) prediction models and ARDS phenotype clusters. We derived 'Patient Context Vectors' representing patient-specific contextual ARDS risk factors, utilizing deep-learning techniques on ICD and free-text clinical notes data. The Patient Context Vectors were combined with structured data from the first 24 hours of admission, such as vital signs and lab results, to build an ARDS patient prediction model and an ARDS patient mortality prediction model achieving AUC of 90.16 and 81.01 respectively. The ability of Patient Context Vectors to summarize patients' medical history and current conditions is also demonstrated by the automatic clustering of ARDS patients into clinically meaningful phenotypes based on comorbidities, patient history, and presenting conditions. To our knowledge, this is the first study to successfully combine free-text and structured data, without any manual patient risk factor curation, to build real-time ARDS prediction models.

Introduction

Critical care Clinical Decision Support (CDS) systems aim at the early identification and timely treatment of rapidly progressive, life-threatening conditions. In particular, ARDS (Acute Respiratory Distress Syndrome) is a significant cause of morbidity and mortality in the USA and worldwide1,2. Early recognition and evidence-based management of ARDS can limit the propagation of lung injury and significantly improve patient outcomes3.

To date, there exists no accurate and reliable way to anticipate which patients, presenting with respiratory distress, are likely to develop ARDS. Numerous prediction scores have been developed to assess ARDS prognosis and risk of death, such as Lung Injury Score (LIS)4, Lung Injury Prediction Score (LIPS)5, APPS (Age, Plateau, PaO2/FiO2 Score)6, Early Acute Lung Injury (EALI)7, and Modified ARDS Prediction Score (MAPS)8. Using a consensus process, a panel of experts convened in 2011 to develop the Berlin definition, focusing on addressing a number of limitations of prior definitions9. Still, the predictive validities of these tools and definitions have proven to be moderate, for example, as measured by area under receiver operating curve (AUC).

The difficulty in analyzing and predicting ARDS outcomes stems from the fact that this is both rare, and, at the same time, highly heterogeneous condition10. ARDS involves the interaction of multiple risk factors, past history, and current conditions, signs, and symptoms. Hospital alert systems typically rely on highly sensitive screening of structured data, such as vital signs and lab results, which, in the case of such rare conditions, are often associated with false clinical alarms resulting in “alarm fatigue”11.

EMR data depends on what the clinician deems necessary to measure and record in the act of caring for the patient. EMR data is typically entered for the purposes of clinical documentation and billing12, and thus not centered around the needs of real-time surveillance-based CDS systems. The combined physician care related variables and underlying patient-related contextual factors needed for a reliable ARDS risk evaluation are typically dispersed across the patient EMR record, and available at different times throughout the patient stay. Patient demographics, past medical and visit history, chronic conditions, risk factors, current signs and symptoms can be found in diverse combinations of structured elements and clinical notes (e.g. nursing notes, radiology reports, etc.), that record diagnosis and procedure codes, vital signs, lab orders and results, ventilation parameters, etc. The challenge for real-time surveillance-based CDS systems is accommodating for the variability and the availability of real-time electronic data and enabling accurate contextual interpretation of real-time patient data.

In this work, we utilize all available EMR patient information, in the form of structured data and free-text, for real-time predictive modeling. While our experiments are focused on identifying ARDS cases, the described method is
applicable to a variety of disease surveillance CDS use cases, needing information dispersed across the EMR patient record.

A second goal of this study is to utilize the combination of clinician knowledge and experience, and a data-driven approach to identify ARDS patients' phenotypes and risk factors, acknowledging the need for targeted personalized treatments reflecting differences in treatment outcomes across patient subtypes.\textsuperscript{13–15}

**Dataset**

Clinical encounter data of adult patients were extracted from the MIMIC3 Intensive Care Unit (ICU) database.\textsuperscript{16} MIMIC3 consists of retrospective ICU encounter data of patients admitted into Beth Israel Deaconess Medical Center from 2001 to 2012. Included ICUs are medical, surgical, trauma-surgical, coronary, cardiac surgery recovery, and medical/surgical care units. MIMIC3 includes time series data recorded in the EMR during encounters (e.g. vital signs/diagnostic laboratory results, free text clinical notes, medications, procedures, etc.). The dataset contains data associated with over 58,000 ICU visits, including over 2 million free-text clinical notes and over 650,000 diagnosis codes.

For this study, in accordance with previous literature,\textsuperscript{17} we identified ARDS for adult patients older than 18 years with ICD-9 codes for severe acute respiratory failure and use of continuous invasive mechanical ventilation, excluding those with codes for acute asthma, COPD and CHF exacerbations.\textsuperscript{1} This resulted in 4,624 ARDS cases from a total of 48,399 adult ICU admissions. The ICU mortality rate in this population was approximately 59%, somewhat higher than expected for ARDS\textsuperscript{18}, suggesting that the algorithm used is capturing the most severe cases of ARDS, and thus introducing some level noise with possibly containing true ARDS cases marked as negative examples.

Our ARDS predictive model utilized data in the form of free-text clinical notes, ICD codes, and structured physiological and ventilator data. The structured data included in this analysis consists of anion gap (aniongap), albumin, bands, bicarbonate, bilirubin, creatine, chloride, glucose, hematocrit, hemoglobin, lactate, platelet, potassium, partial thromboplastin time (ptt), international normalized ratio (inr), prothrombin time (pt), sodium, bun, white blood cell count (wbc), heart rate (heartrate), systolic blood pressure (sysbd), diastolic blood pressure (diabs), mean blood pressure (meanbp), respiratory rate (resperate), body temperature (tempc), peripheral capillary oxygen saturation (spo2), body mass index (bmi), gender, age, urine output (urine1). All variables are included as min, max, and mean values and are measured over the first 24 hours of ICU admission. The first 24 hour timeframe was chosen, as it has been reported that ARDS develops at a median of 30 hours after hospital admission.\textsuperscript{19} Thus, a 24-hour window provides for the gathering of enough structured data, while at the same time is early enough for real-time CDS.

**Descriptive Analytics**

ARDS patient characteristics and risk factors were first gathered with the help of experienced clinicians. Expert knowledge was gathered in the form of Concept Maps (Cmaps).\textsuperscript{20} Cmaps is a tool developed at the Institute for Human and Machine Cognition (IHMC) that enables collaborative knowledge creation, in the form of concepts, relations, and ontologies, with links to external resources and publications. A snippet of the developed ARDS Cmap developed by our clinical research team is shown in Figure 1.

For example, clinicians coded ARDS precipitant causes include sepsis, aspiration, traumatic injuries, burns, and drugs, including illicit drugs, such as cocaine, heroin, or prescription drugs, such as chemotherapeutic agents, etc.

Initially, the ARDS Cmap was used as a screening rule engine. In addition, the Cmap was used to identify risk factors that were later used in predictive models. Although rule engines based on Cmaps are evidence-based and tend to be highly sensitive, they tend to perform with subpar specificity (e.g. the Berlin definition achieved an AUC of 0.577). Such rule engines are also highly sensitive to missing data. In contrast, data-driven and Machine Learning algorithms have the potential to improve on rule engines from training on large datasets, learning from a variety of clinical data and response patterns, and are able to handle missing data. However, for machine learning to be effective in predicting highly heterogeneous conditions such as ARDS, training data requires both high precision labeling and the identification of features with adequate numbers of samples needed to separate classifications of ARDS classes and subclasses/phenotypes from non-ARDS patients.

\textsuperscript{1}Inclusion ICD9 Codes: 51881, 51882, 51884, 51851, 51852, 51853, 5184, 5187, 78552, 99592, 9670, 9671, 9672; Exclusion ICD9 Codes: 49391, 49392, 49322, 4280.
Figure 1: ARDS Cmap: Clinician-coded representation of ARDS patient characteristics and risk factors.

ICD Embeddings and Patient Vectors

We then looked for a data-driven approach to provide additional insight into ARDS patient characteristics, risk factors, and phenotypes.

Intuitively, even without any additional patient EMR data, clinicians viewing properly coded patient diagnosis codes (e.g. ICD codes in a problem list) are typically able to create a mental summary of the overall patient condition, including medical history, risk factors, presenting conditions. ICD codes are used to describe both current diagnoses (e.g. Pneumonia, unspecified organism: ICD9 486), but also a variety of additional patient information. For example, ICD codes can describe ARDS risk factors, such as patient’s history and chronic conditions (e.g. Chronic kidney disease; Personal history of malignant neoplasm; etc.); information regarding past and current treatments and procedures (e.g. Infection due to other bariatric procedure). In some cases, ICD codes contain information such as the patient age group and/or susceptibilities (e.g. Sepsis of newborn; Elderly multigravida); expected outcome (Encounter for palliative care); patient social history (e.g. Adult emotional/psychological abuse; Cocaine dependence); the reason for the visit, (e.g. Railway accidents; Motor Vehicle accidents).

Using ICD codes for statistical analysis and predictive models, however, poses a series of challenges. Patient ICD codes in EMRs tend to be sparse. There are numerous ICD codes (around 15,000 ICD9 codes and around 68,000 ICD10 codes), with only a very small subset of these applicable to a particular patient (e.g. MIMIC3 admissions have an average of 11 ICD codes). ICD codes also tend to co-occur and overlap. In addition, ICD coding can be, in some cases, subjective and dependent on numerous external factors.

However, the concurrence and mutual information of ICD codes over large data repositories can be utilized. For example, the fact that Pneumonia ICD codes are often accompanied with ICD codes describing Cough, Fever, Pleural effusion, etc. can be utilized to generate vector representations of ICD codes. Inspired by deep learning representation, such as word embeddings, it has been suggested that this medical code co-occurrence can be exploited to generate low-dimensional representations of ICD codes that may facilitate EMR data-based exploratory analysis and predictive modeling.

In this study, we utilized available MIMIC3 patient data to generate the ICD embeddings following the approach of Choi et al. In our approach, we generated a low-dimensional representation of the patient history, symptoms, risk factors, diagnosis, etc., by averaging the patient ICD code embeddings. We refer to this representation of the patient’s medical history and clinical condition as Patient Context Vectors.

To generate ARDS patients groups sharing similar characteristics and risk factors, we then clustered the ARDS Patient Context Vectors via k-means clustering. Clinical review of the generated ARDS clusters determined the optimal number of clusters to be 10.
The Patient Context Vectors were able to clearly separate ARDS patient risk factors and conditions into clinically valid categories, such as Malignancy or Chronic Hepatic Disease. Figure 2 shows the frequency of patients in various clusters, sorted left-to-right according to mortality rate. Table 1 lists the 15 most representative ICD code descriptions (cluster centroids) for the 10 ARDS clusters. The cluster description was provided by clinicians reviewing the corresponding cluster centroids.

![Figure 2: Frequency and mortality rate of MIMIC3 ARDS patient clusters formed by clustering of averaged ICD embeddings.](image)

Interestingly, the manually clinician-curated Cmap appears to overlap to a large extent with the automatically derived ARDS patient clusters. For example, clinicians listed chemotherapeutic agents as a distinct risk factor for ARDS. Our data-driven analysis showed a distinct cluster of ARDS patients with ICD codes describing malignancy. Further analysis could delineate if this is a direct relationship between chemotherapeutic agents and the development of ARDS, or rather reflects an indirect relationship. For example, perhaps chemotherapy-induced immunosuppression acts as a risk factor for sepsis which, in turn, acts as a risk factor for ARDS.

Furthermore, the mortality associated with different clusters was also consistent with clinician experience. Clinicians recognize that patients with advanced malignancy may develop severe infections and ARDS as a final common pathway in their advanced disease. In this context, ARDS may represent a manifestation of their advanced underlying disease, and therapies directed at ARDS and its precipitant may not significantly impact mortality. Indeed, the cluster of malignancy-associated ARDS had a very high mortality rate of 90%.

In contrast, patients who develop ARDS as a result of trauma typically were well enough to be engaged in the activity leading to trauma, and thus ARDS may truly be the primary disease as opposed to a symptom of advanced underlying disease. The likelihood of mortality in this group may be more significantly influenced by the treatments targeted at ARDS. In our analysis, the cluster of trauma associated-ARDS had a significantly lower mortality rate of 30% compared to other clusters.

**Predictive Analytics**

Clustering and manual evaluation by clinicians of Patient Context Vectors proved to be a useful tool in summarizing the overall patient condition, risk factors and history. Real-time CDS systems, however, might not have access to the full set of the patient ICD codes as they might be entered in the EMR system at a later stage. It has been observed that clinical notes, specifically nursing and physician notes, typically contain all of the information available from ICD codes. Furthermore, while past medical history and presenting conditions might not always be ICD-coded, they are typically available in the form of free-text notes. In previous work, we have successfully utilized free-text for predicting Patient Context Vectors, in the case of missing ICD codes. Additionally, we were able to successfully combine information available in ICD codes and in nursing notes and produce an average Patient Context Vector.

A word-level Convolutional Neural Network (CNN) was trained to predict from free-text notes the Patient Context Vector (averaged ICD code embedding). Structured data, in the form of vital signs and lab results, was then combined with the predicted Patient Context Vector and utilized in a machine learning model trained to predict ARDS patients.
Table 1: The top 15 most representative ICD codes for various clusters, based on cosine similarity to the cluster centroid.

Figure 3 summarizes the system workflow during prediction time.

The ARDS prediction model was trained utilizing structured data from the first 24 hours of admission, as described above in the Dataset Section. In addition, the first half of the available patient nursing notes and the first half of entered ICD codes were used to produce Patient Context Vectors of size 50. The patient ICD codes were used to look...
Figure 3: Real-time ARDS prediction workflow. Nursing notes available at prediction time are used to predict Patient Context Vectors. ICD codes available at prediction time are also converted to Patient Context Vectors by averaging ICD code embeddings. Patient Context Vectors are used together with structured EMR data to predict the patient ARDS status.

<table>
<thead>
<tr>
<th>ARDS Prediction</th>
<th>Features</th>
<th>AUC</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>79.74</td>
<td>29.77</td>
<td>67.80</td>
<td>37.17</td>
<td></td>
</tr>
<tr>
<td>Baseline + first half of notes/ICD</td>
<td>90.16</td>
<td>49.46</td>
<td>78.85</td>
<td>48.63</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARDS Mortality Prediction</th>
<th>Features</th>
<th>AUC</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>78.26</td>
<td>69.19</td>
<td>92.60</td>
<td>79.20</td>
<td></td>
</tr>
<tr>
<td>Baseline + first half of notes/ICD</td>
<td>82.11</td>
<td>73.66</td>
<td>89.99</td>
<td>81.01</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: 10-fold cross-validation GBM results of predicting ARDS patients and predicting mortality among ARDS patients. P=Precision, R=Recall, F1= F1-score for the positive class. The Baseline set of features consists of vital signs, lab results, Glasgow Coma Scale score, gender and age, in the form of structured data. "Baseline + first half of notes/ICD" includes also the average of the first half of entered visit ICD codes embeddings, and Patient Context Vectors predicted from the first half of the visit nursing notes.

up the corresponding ICD code embedding and averaged. Each nursing note was used to predict the Patient Context Vector via the trained word-level CNN. All Patient Context Vectors were then averaged and used in addition to the structured MIMIC3 data. As the MIMIC3 dataset ICD codes lack timestamps, we were unable to identify ICD codes available during the first 24 hours of admission. The MIMIC3 ICD Codes, however, are ordered, and we used the first half of the ICD codes, and the first half of notes, as an approximation of data available early in the patient stay.

A Gradient Boosting Machine (GBM) model\textsuperscript{31,32} was used to predict ARDS patients from the total population of adult patients. A GBM model was also used to predict the mortality among all ARDS patients. Table 2 shows the result from the experiments. All results were produced via 10-fold cross validation.

In both prediction models, the inclusion of Patient Context Vectors significantly increased the overall model performance. Intuitively, the patient medical history and overall clinical condition are important predictive factors. Results demonstrate that Patient Context Vectors can be successfully utilized to represent additional knowledge of a patient’s condition. The importance of the Patient Context Vector is also demonstrated by the scaled GBM variable importance shown in Figure 4. Together with critical vital signs measurements, various Patient Context Vector dimensions (shown with prefix embedding) play an important role in predicting the patient’s ARDS outcome. A known limitation of
low-dimensionality representations is the lack of interpretability of individual dimensions (e.g. embedding 39 lacks the interpretability of systolic blood pressure or temperature). Future work will focus on interpretability-imparted patient context vector embeddings, amenable to clinical interpretation.

In terms of practical application, the proposed system can be used in addition to existing high recall/low precision hospital alert systems, and used to prioritize alerts, mitigating the effects of alert fatigue. Furthermore, the imperfect and noisy nature of the automatically created dataset is likely resulting in over-pessimistic evaluation. It has been shown that ML classification algorithms are able to achieve high performance at relatively high levels of noise and that ML models generated from noisy datasets perform significantly better when evaluated on clean test sets (10 to 30% classification accuracy improvement at high training set noise levels)\(^33\)–\(^35\). Future work will focus on creating a clean, clinician-reviewed ARDS test dataset for more precise evaluation of the proposed approach.

**Related Work**

Numerous prediction scores have been developed to assess ARDS prognosis. Gajic et al.\(^5\) developed a Lung Injury Prediction Score (LIPS) formula including predisposing conditions, such as sepsis, shock, pneumonia, alcohol abuse, chemotherapy, FIO2 and respiratory rate measures, found useful in predicting ARDS and mortality in surgical critical care patients\(^36\). Interestingly, over 70\% of score points associated with LIPS score calculation are based on patient context data rather than vials, labs, symptoms.

Other tools, such as Villar et al.\(^6\) base their ARDS prediction score on Age, PaO2/FIO2, and Plateau pressure score. Levitt et al.\(^7\) developed an Early Acute Lung Injury (EALI) score including risk factors, respiratory rate, and oxygen requirement. Xie et al.\(^8\) developed a modified ARDS prediction score (MAPS) based on a hand-crafted set of risk factors, risk modifiers, vital signs, etc.

A number of studies focus on predicting mortality among ARDS patients. For example, similar to our cluster findings, Hyers\(^37\) reports that patients who develop bacterial sepsis and multiple organ dysfunction are at high risk of dying and patients who develop ARDS from trauma or other noninfectious causes have a better prognosis. Navarrete-Navarro et al.\(^38\) report that mortality among ARDS patients correlates with the PaO2/FIO2 ratio on the 3rd day of ARDS, the APACHE III score, and the development of multiple organ system failure. Timmons et al.\(^39\) studied children with ARDS and found significant differences between survivors and nonsurvivors who develop ARDS from trauma or other noninfectious causes have a better prognosis. Villar et al.\(^40\) built an ARDS mortality prediction model based on tertiles of patient age, plateau airway pressure, and PaO2/FIO2 at the time the patient meets ARDS criteria. Similarly, Spicer et al.\(^41\) studied pediatric ARDS patients and determined that oxygenation index and hematopoietic stem cell transplant / cancer history can be used on Day 1 or Day 3 of ARDS to predict hospital mortality without the need for more complex models.

Unlike the current work, the described previous studies utilized only structured patient data, and ICD codes/risk factors, when used, consisted of manually crafted lists.

In a broader context, a large volume of literature on combining structured and free-text EMR data apply Medical Concept detection on the free-text notes for manually curated list of risk factors and other disease-relevant medical
concepts. Ford et al. present a review of various approaches to Medical Concept detection from free-text notes for the purpose of detecting cases of a clinical condition, often in conjunction with structured data.

More recently, deep learning has been used to utilize free-text and structured EMR data. Shickel et al. present a survey of various deep learning techniques. Miotto et al. build a Deep Patient representation in an unsupervised manner via denoising autoencoders, however, similar to previous approaches they first pre-process the free-text notes by extracting medical concepts with an off-the-shelf tool. Various studies use deep learning techniques to generate low-dimensional representations of diagnosis codes and patients utilizing structured data (diagnosis codes, medications, and procedures). Unlike previous work, we combine free-text and structured EMR data for obtaining low-dimensional patient representations, without the use medical concept detection.

Conclusion
This work demonstrates the utility of deep learning techniques to summarize a patient’s medical history, risk factors, comorbidities, and current signs and symptoms in the form of Patient Context Vectors. Automatically generated ARDS patients clusters agree with manually curated clinician knowledge and provide additional insight into the complexities and risk factors associated with ARDS. More importantly, Patient Context Vectors, derived from available ICD codes and nursing notes, can be easily combined with structured EMR data to build real-time ARDS CDS tools, with potential to improve patient outcomes and reduce mortality among ARDS patients.

Acknowledgements
Research reported in this publication was supported by a NIH SBIR award to CTA by NIH National Heart, Lung, and Blood Institute, of the National Institutes of Health under award number 1R43HL135909-01A1.

References


Non-Negative Matrix Factorization for Drug Repositioning: Experiments with the repoDB Dataset

Gokhan Bakal, M.S.1, Halil Kilicoglu, Ph.D.2, Ramakanth Kavuluru, Ph.D.1

1University of Kentucky, Lexington, KY; 2National Library of Medicine, Bethesda, MD.

Abstract
Computational methods for drug repositioning are gaining mainstream attention with the availability of experimental gene expression datasets and manually curated relational information in knowledge bases. When building repurposing tools, a fundamental limitation is the lack of gold standard datasets that contain realistic true negative examples of drug–disease pairs that were shown to be non-indications. To address this gap, the repoDB dataset was created in 2017 as a first of its kind realistic resource to benchmark drug repositioning methods — its positive examples are drawn from FDA approved indications and negatives examples are derived from failed clinical trials. In this paper, we present the first effort for repositioning that directly tests against repoDB instances. By using hand-curated drug–disease indications from the UMLS Metathesaurus and automatically extracted relations from the SemMedDB database, we employ non-negative matrix factorization (NMF) methods to recover repoDB positive indications. Among recoverable approved indications, our NMF methods achieve 96% recall with 80% precision providing further evidence that hand-curated knowledge and matrix completion methods can be exploited for hypothesis generation.

1 INTRODUCTION
Repositioning previously approved drugs for new indications has become highly desirable in the biomedical and pharmaceutical research enterprises given expected time/cost reductions in identifying new treatment options. With recent estimates putting new drug development R&D costs over $2.5 billion per drug1, repurposing has been gaining mainstream attention over the past five years. With previously approved drugs already passing the required safety tests for use in humans, the cost of repositioning is expected to be substantially lower compared with starting from a blank slate. In 2011, the National Library of Medicine (NLM) introduced drug repositioning (DR) as a new Medical Subject Heading (MeSH term) and as of now there are 1,161 articles tagged with it dating back to a single article from 2009. Almost 85% of these articles are published in the last five years indicating the sudden and deserved surge of interest in this area. Physicians with deep understanding of both the mechanisms of action for drugs and disease characteristics may be able to recommend off-label use2 in an ad hoc manner. However, this does not constitute FDA approved recommendation for specific new indication(s) of drugs for use in designated groups of patients. As such, DR (via FDA approval) for new indications has significant potential to impact care at a broad scale and might also result in lowered costs for patients.

1.1. Computational drug repositioning. Computational drug repositioning (CDR) is the use of informatics and high performance computing methods to prioritize candidates for new indications. With the simultaneous excitement surrounding biomedical data science and the explosive growth of publicly shared datasets, CDR methods are on the rise in the scientific community. A class of such methods exploits the notion of “similarity” between different entities involved in disease and therapeutic mechanisms. For example, shared traits among drugs including chemical structures, molecular activities, and side effects may be used to define a feature vector to represent a drug. Likewise, similarities can also be established between diseases based on established gene–disease associations or graph based proximity in disease ontologies. Zhang et al.4 provide a unified framework that exploits these similarities for CDR. Another direction of CDR is exploiting available large-scale genomic data sources. For instance, Dudley et al.5 utilized drug-gene expression signatures to discover a potential new drug for the inflammatory bowel disease. Topological analyses of drug–target networks and target-involved pathways are another mode of identifying potential new indications.6 Text mining programs that extract different relations from text using natural language processing (NLP) and literature based discovery approaches that build on such relations are also being employed for CDR. A more detailed treatise of CDR methods is available in a recent survey3.

1.2. CDR method assessment. Although CDR is gaining prominence, evaluating CDR methods can be tricky given the lack of datasets that are tailored for it. Specifically, from our literature review we were able to identify very few
standardized dataset\cite{tesinat2018metathesaurus} that are uniformly used across efforts for benchmarking purposes. Furthermore, the datasets used in prior efforts have a serious shortcoming — they only contain positive drug–disease indication pairs; and hence prior efforts assume that all other combinations are negatives, which is unreasonable and potentially rules out novel repositioning predictions as false cases. Brown and Patel\cite{brown2015treatment} highlight this shortcoming and propose a new gold standard database called repoDB for CDR method benchmarking. repoDB draws approved indications from the DrugCentral\cite{drugcentral} database and failed indications from the American Association of Clinical Trials Database (the ‘AACT Database’\cite{aaact}), which is a structured version of information from NLM’s ClinicalTrials.gov service. Given failed indications are part of the dataset, one can directly assess CDR methods with vetted indications and non-indications from repoDB. Since its introduction in 2017, however, we are not aware of any CDR efforts evaluating against repoDB.

In this paper, we present the first CDR attempt that directly tests against repoDB instances. First, a partially observed matrix is built using drug–disease treatment relations drawn from the UMLS Metathesaurus\cite{metathesaurus} and those extracted using automated NLP methods and made available by the NLM as part of the SemMedDB database\cite{semmed}. Next, this matrix is completed by filling unobserved cells via non-negative matrix factorization (NMF) to elicit new indications. Our method uses a small portion of repoDB as a validation dataset and uses the bulk of it for testing purposes.

\section{Materials and Methods}

\subsection{Datasets}

In this section, we describe the data sources from which we derive our training and testing examples. The UMLS\cite{umls} and SemMedDB\cite{semmed} are our essential data resources for training instances while repoDB is our resource for the test examples. We use the terms “training” and “testing” to emphasize that this is still a (weakly) supervised method where the training instances are simply drawn from external resources both manually curated (UMLS) and automatically extracted (SemMedDB). We will briefly describe each data source in the following subsections.

**UMLS Metathesaurus.** UMLS is a longstanding terminological resource that integrates over 160 different vocabularies updated every year by the NLM. The Metathesaurus portion of UMLS aggregates equivalent concepts across multiple vocabularies and assigns to each unique concept a concept unique identifiers (CUI). Besides synonymous names for each concept, there are also inter-concept relations sourced from the original vocabularies. We obtained treatment relations from the MRREL table\cite{mrrel} in UMLS Metathesaurus\cite{umls} version 2017AB. A total of 43,898 such relations are part of our UMLS training dataset.

**SemMedDB – Semantic Medline Database.** SemMedDB is a repository of (subject, predicate, object) triples called semantic predications extracted by a rule-based NLP tool SemRep\cite{semrep} developed by the NLM. SemMedDB is built by running SemRep over all available PubMed citations (over 27 million) where the subject/object entities are normalized to UMLS CUIs. Likewise, the predicate is mapped to a relation type from the UMLS semantic network\cite{umls}. Given a predication can be extracted from multiple sentences, we also have frequency information (number of unique sentences containing it) for each SemMedDB triple. For our experiments, we curated treatment predications (triples where predicate = TREATS) in SemMedDB as additional training examples. As SemRep’s precision is around 75\%\cite{semrep}, we only collected predications which have been extracted at least twice, thrice, and five times to include them in the training set in various configurations (more later). Hence, we were able to obtain three different treatment predication sets of 55,349, 34,802 and 19,977 triples for the frequencies of 2, 3, and 5 respectively as long as they are not occurring in test sets from repoDB.

**The repoDB database.** As indicated in Section\cite{tesinat2018metathesaurus} instances in repoDB come from DrugCentral\cite{drugcentral} and ClinicalTrials.gov\cite{clinicaltrials} resources. It has a total of 6,677 approved and 3,885 failed drug–disease pairs. After removing the duplicates and the ones which appear in UMLS (given UMLS pairs will be part of the training dataset), we were left with 6,218 approved and 2,852 failed pairs. After removing pairs associated with drugs for which there is not even a single positive pair from UMLS/SemMedDB, we are left with 5,172 approved treatments (ATs) and 2,244 failed indications (FIs). This aligns with the nature of CDR to some extent — if we do not even have a single occurrence of a drug treating some disease, we may not be able to repurpose it for other conditions. This is also an inherent limitation of the matrix completion method we propose to use; if the row corresponding to a drug in the drug–disease matrix is empty, matrix completion methods cannot fill that row and hence it is impossible to come up with new indications for it (more later).

\*Specifically, these are the relations where the RELA field in MRREL table is equal to one of these four types: “treats”, “may_treat”, “treated_by”, and “may_be_treated_by”
**Generation of randomly selected negative examples.** The repoDB examples are vetted ATs and FIs, identified based on clinical trials. We also wanted to build a separate dataset of random indications (RIs) which satisfy domain/range constraints for subjects/objects for `treats` predicate. The purpose is to see if our method would have a relatively easier or harder time when dealing with these when compared with FIs from repoDB. In the past we have generated such a dataset for a slightly different task. Basically, these RI examples are created by the following steps.

- Each concept in UMLS has at least one semantic type that represents a class membership. Furthermore, every predicate in the UMLS semantic network has a set of domain/range semantic type constraints defined by the NLM based on domain expert knowledge. Based on the allowable semantic type combination for the `treats` predicate, we randomly select pairs that satisfy the domain/range constraints.
- For the set of pairs selected using the previous step, we simply remove the pairs which appear as treatment relations either in UMLS or SemMedDB. Thus, we ensure that selected pairs do not occur in our training set.

The given steps above pick fairly hard-to-predict potentially negative examples because they satisfy the domain/range constraints and are not present in either UMLS or SemMedDB databases. Ultimately, we obtained 3,318 examples to be used as the RI test set for the matrix completion methods.

### 2.2. Methods

In this section we present the NMF based matrix completion method along with our approach to configure it with different input matrices from external data sources.

#### Matrix completion through NMF

Matrix completion is the process of filling missing entries in a partially observed matrix. These partially observed matrices arise in many real world scenarios especially in recommender systems where preferences of people are encoded. A matrix with customers as rows and and products (e.g., movies, books) as columns is the typical setup. Given information about their prior ratings or product purchases represented as 1s in the corresponding cells, matrix completion would identify what other cells ought to be 1s — which other products would a customer likely enjoy given what they already liked. In a completely random world, there is no way to guess the new 1s. However, assuming the matrix has a much smaller rank than \( \min(m,n) \) for the \( m \times n \) matrix, we can use non-negative matrix factorization (NMF) to come up with a low-rank approximation to the original matrix with \([0,1]\) non-zero entries in blank cells, leading to potential new recommendations. This low-rank assumption is based on the intuition that there are latent themes/trait in user preferences and a typical user’s preferences are not distributed truly randomly across the product space. A similar strategy is also employed in information retrieval for latent semantic indexing for computing document similarity through dimensionality reduction.

One can now see that the CDR problem can be modeled similarly where the training treatment relations can be used to partially fill the drug–disease matrix, with NMF filling empty cells with non-zero values pointing to potential new indications. Since this is an approximation process, the new values in empty cells will be non-zero but generally not exactly 1. Thresholding based on a validation dataset can be used to glean indications if a particular cell’s value crosses the threshold. The intuition here is also to exploit potential latent themes where groups of drugs sharing certain characteristics (e.g., mechanism of action) may treat clusters of conditions with similar traits (e.g., symptoms). Given we do not know what the myriad latent themes may be, we assume a certain number of them are present — the chosen low rank — and proceed with NMF for matrix completion. Thus, given the partially observed \( m \times n \) drug–disease matrix \( X \) with \( m \) drugs and \( n \) diseases, we will approximate it as

\[
X \approx W_{mxn} \times H_{kxn} = \hat{X}_{mxn}, \tag{1}
\]

where \( W \) and \( H \) are the factors with rows of \( W \) representing \( k \)-dimensional drug vectors and columns of \( H \) encoding \( k \)-dimensional disease vectors under the assumption that \( X \) has rank \( k \ll \min(m,n) \). The product \( \hat{X} = WH \) approximates \( X \) helping us glean new non-zero values hinting at new indications, while the rows of \( W \) and columns of \( H \) can be used to compute drug and disease similarities respectively. The objective function to find the best approximation is

\[
\arg \min_{W,H} \|X - WH\| + \beta(\|W\|_2 + \|H\|_2), \tag{2}
\]

where \( W \in \mathbb{R}^{m \times k} \) and \( H \in \mathbb{R}^{k \times n} \) and \( \beta \) is the weight for the regularization penalty term to handle overfitting that corresponds to large norms for \( W \) and \( H \). Next, the construction of the input drug–disease matrix \( X \) is discussed.
Building the input training matrix. The input partially observed matrix was constructed based on treatment relations from UMLS and SemMedDB as indicated in Section 2.1. However, we needed to consider a larger matrix to fill compared with drugs and diseases of positive indications from UMLS Metathesaurus and SemMedDB. Otherwise, the test indications in repoDB cannot be recovered as part of the completion process. For this, we considered all allowed subject/object semantic type constraints for treats predicate (based on SRSTR tables of the UMLS semantic network). Next, we selected all UMLS subject concepts where each has at least one semantic type that belongs to the set of allowed subject types; likewise, we aggregated all UMLS concepts where each has at least one semantic type that is from allowable object types. Based on this, 538,710 subject entities and 314,707 object entities were obtained for the input matrix. However, most subjects do not have a treatment relation with any disease as observed in the Metathesaurus training dataset. Hence, we removed all zero rows (hence the corresponding drugs or treatment agents) and retained only rows that are known to treat \( \geq 1 \) disease in the columns to exploit the shared therapeutic context among the drugs. After this pruning process, we were left with 10,188 drugs (or treatment agents) and 314,707 objects to build our input matrix to be partially filled (from training datasets) and subsequently completed via NMF.

To populate the input matrix with training relations, we have Metathesaurus treatment relations and three different sets of semantic predicates derived from SemMedDB based on extraction frequencies. We have a configuration where training relations are entered in the tables as 1s, which we call the “binary matrix factorization (BMF)” model. In addition to BMF, we have another configuration using SemMedDB treatment predicates with their extraction frequency counts. We term this as the “count matrix factorization (CMF)” model to evaluate the performance when counts are used instead of Boolean indicators.

Finally, as part of the input we have many unallowed input matrix cells (5,531,386 in total) that cannot be 1s because

---

1As an example, (Pharmacologic Substance, Disease or Syndrome) is a popular semantic type combination allowed for treatment relations. (Antibiotic, Disease or Syndrome) and (Therapeutic or Preventive Procedure, Congenital Abnormality) are less common allowed type combinations. Overall, there are 56 different allowed type combinations for treatment relations as per UMLS. There are other additional allowed types that NLM has incorporated as part of the schema design for SemMedDB and those are included for this effort as allowed combinations.

2Note that not all objects are diseases per se, some maybe symptoms and at times different patient groups. For example, the UMLS CUI C4316221 refers to “Patients with a diagnosis or past history of total colectomy or colorectal cancer” and can be an object of some treatment relations. To capture all latent themes we end up using all subjects and objects of treatment relations as part of the input matrix. However, for evaluation purposes, after NMF, we only look at those cells (in \( X \) from Eq. (1)) corresponding to approved and failed indication pairs in repoDB.
the corresponding subject–object pairs do not satisfy the domain/range semantic type constraints. For example, consider this unallowed type combination: \((\text{Drug Delivery Device, Patient or Disabled Group})\). Although \text{Drug Delivery Device} is an allowed subject type for some other type combination(s) and \text{Patient or Disabled Group} is an allowed object type for a different combination, this particular coupling is not allowed. But cells corresponding to this unallowed combination exist in the input matrix given the matrix was built with all allowed subjects and objects as rows and columns, respectively. Thus, these cells corresponding to entity pairs that satisfy this type combination must be designated as unallowed. To avoid such predictions, we assign 0s to the corresponding cells of the input matrix. This is another way to further constrain the factorization process to approximate both positive and unallowed cases while estimating the unobserved cells. We experimented with several input matrices with different numbers of unallowed cases to observe their influence on the prediction task. The levels of these unallowed examples are set to 0% (no unallowed examples), 25%, 50%, 75%, and 100% (all of them) in the input matrix. The overall framework of our approach is demonstrated in Figure 1.

**NMF experimental configurations.** We note that repoDB drugs/diseases for ATs and FIs are already mapped to UMLS CUIs by its creators. Hence, the matrix constructed and completed as described in Section 2.2 naturally suffices for the CDR task. When training, experiments were conducted with singular value decomposition (SVD) to identify a \(k\) value (to be used for the dimensionality in Eq. (1)) that minimizes the mean squared error (MSE) for the cells that are already filled in the training matrix. Since there were not any noticeable differences between MSE values with \(k = 50, 100, 500\), we chose \(k = 50\) for our further matrix completion experiments. Thus results are reported for \(k = 50\) for all experiments. The regularization parameter \(\beta\) was left at the default value (0.1) because tuning it did not yield any apparent gains. To carry out the optimization in Eq. (2), the open source MF library LIBMF\(^\text{23}\) was used for incomplete matrix approximation. LIBMF is an efficient stochastic gradient descent based software package that runs parallel on multiple cores in a shared-memory environment.

3 RESULTS

We assess NMF results from two different perspectives. First we look the actual NMF scores produced for ATs, FIs, and RIs we created as part of Section 2.1. Subsequently, we observe how these NMF scores can be used to come up with precision, recall, and F-score based on ATs and FIs from repoDB.

3.1. NMF scores for repoDB pairs. In Table 1 we show mean NMF scores for ATs and FIs in repoDB and the RIs we generated. The scores are real-valued numbers with which NMF fills unobserved cells as part of the training process for various configurations. Configurations differ from each other with respect to what is included in the input matrix and the proportions of unallowed pairs included as 0s. Due to the optimization in Eq. (2) and encoding of positive cases as 1s, the higher the value estimated for an unobserved cell, the stronger the plausibility of treatment relationship for the corresponding drug–disease pair. We make the following important observations from Table 1:

- Only adding positive training examples to the input matrix and leaving unallowed examples as unobserved (the 0% column) leads to catastrophically bad results where the FIs and RIs are scoring higher. Hence we do not discuss any results going forward where the unallowed examples are left as unobserved. Adding additional unallowed examples (all other columns) as 0s in the input matrix shows more realistic scores following a clear pattern where the ATs score higher than FIs, which fare better than RIs. This confirms a few things: (a). The 0s inserted to account for unallowed cases are providing enough signal to guide the optimization process to distinguish between more plausible indications from random ones (which are mostly useless). (b). NMF based completion is able to score ATs better than FIs in repoDB where the mean AT score is 2–3 times higher than that of the FI score, demonstrating its effectiveness. (c). FIs scoring higher than RIs reflects the reality that FIs actually went through the process of clinical trials, which implies researchers felt that those pairs were plausible indications; RIs however are just random pairs of drugs and diseases. (d). Including more unallowed pairs as part of the input quickly decreases the magnitude of the scores (compare the 25% column with the 100% column); however, the relative differences between ATs, FIs, and RIs persist all across the board.
- Adding more training positives from SemMedDB (rows 4–12) increases the absolute values of the scores and also the differences in scores between ATs, FIs, and RIs, but the relative differences are the highest when using just UMLS Methasaurus relations. For example, in the 25% column, the ratio of means of AT and FIs for
Table 1: Mean of the predicted NMF scores of test sets with different configurations

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Data</th>
<th>Test Sets</th>
<th>Portion of included unallowed pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>BMF</td>
<td>UMLS only</td>
<td>Approved treatments</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Failed indications</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random indications</td>
<td>0.995</td>
</tr>
<tr>
<td>BMF</td>
<td>UMLS + SemMedDB</td>
<td>Approved treatments</td>
<td>0.936</td>
</tr>
<tr>
<td>(MinFreq. 5)</td>
<td></td>
<td>Failed indications</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random indications</td>
<td>0.987</td>
</tr>
<tr>
<td>BMF</td>
<td>UMLS + SemMedDB</td>
<td>Approved treatments</td>
<td>0.930</td>
</tr>
<tr>
<td>(MinFreq. 3)</td>
<td></td>
<td>Failed indications</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random indications</td>
<td>0.983</td>
</tr>
<tr>
<td>BMF</td>
<td>UMLS + SemMedDB</td>
<td>Approved treatments</td>
<td>0.927</td>
</tr>
<tr>
<td>(MinFreq. 2)</td>
<td></td>
<td>Failed indications</td>
<td>0.951</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random indications</td>
<td>0.976</td>
</tr>
<tr>
<td>CMF</td>
<td>UMLS + SemMedDB</td>
<td>Approved treatments</td>
<td>34.758</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Failed indications</td>
<td>30.327</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random indications</td>
<td>39.603</td>
</tr>
</tbody>
</table>

“UMLS only” (0.268/0.119 = 2.25) is higher than the corresponding ratio for “UMLS+SemMedDB (MinFreq. 5)” (0.564/0.357=1.58), which stays at a similar level even as additional relations are added (MinFreq values 3 and 2).

- Count based models (instead of the binary models) where frequencies are included in the input matrix appear not as consistent (last three rows) where for the 25% column, we notice RIs scoring higher than FIs.

We computed 95% confidence intervals that showed that the score differences are statistically significant. Here we disclose the intervals for the three rows of “UMLS+SemMedDB (MinFreq. 2)” (of the 25% unallowed cases column): 0.650 ± 0.010 (ATs), 0.413 ± 0.017 (FIs), and 0.195 ± 0.012 (RIs). The intervals do not overlap further confirming the NMF method’s functionality.

3.2. Precision, Recall, and F-score for ATs in repoDB. The NMF score ranges in Section 3.1 demonstrate that, on average, NMF maps ATs, FIs, and RIs to non-overlapping segments on the real number scale with high confidence. However, we still need a way to make repositioning Yes/No decisions at the instance level based on the score generated for a particular drug–disease pair corresponding to an entry in \( \hat{X} \) in Eq. (1). One way to make such a decision is to choose a threshold for the NMF score and assign all pairs with scores above that threshold as new candidates for repositioning. Here we propose to do that by splitting the repoDB ATs and FIs into validation and test sets. We considered 20% of ATs and 20% of FIs as comprising the validation set while the rest are left for the final test.

We identified a threshold based on grid search over the validation dataset optimized for F-score with a small step size of 0.00001 spanning the range \([T_{\min}^v, F_{\max}^v]\) such that

\[
T_{\min}^v = \min(\{\hat{X}_{i,j} : (i,j) \in T^v\}) \quad \text{and} \quad F_{\max}^v = \max(\{\hat{X}_{i,j} : (i,j) \in F^v\}),
\]

where \( \hat{X} \) is the approximation from Eq. (1) and \( T^v \) and \( F^v \) represent the validation datasets for ATs and FIs, respectively. This range was chosen based on the observation on the validation dataset that \( T_{\min}^v \) is smaller than \( F_{\max}^v \) (so there were some AT scores that were less than other FI scores). Hence choosing \( T_{\min}^v \) as the threshold corresponds to 100% recall and selecting \( F_{\max}^v \) leads to 100% precision. Thus by limiting the grid search to the threshold range \([T_{\min}^v, F_{\max}^v]\), we are exploring the space of compromise between perfect precision and perfect recall.

\[\text{This translates to 4138 ATs and 1795 FIs in the test set and 1034 ATs and 449 FIs in the validation dataset — numbers computed based on the original repoDB counts from Section 2.1.}\]
Once a threshold is chosen to make instance level decisions for test examples, it is straightforward to assess the performance of the method using traditional measures such as precision, recall, and F-score. Thus, in Table 2, we report the performance results for the BMF models for different configurations of the input matrix and different levels of included unallowed examples. The first observation is that the thresholds selected are all very close to zero indicating that boundary case NMF scores were close to zero for ATs across all configurations. The thresholds are identical for configurations with SemMedDB examples but change slightly for UMLS-Only case. We notice that the best F-score of 0.895 (first row, last column) is obtained for UMLS-Only input matrix with 25% unallowed example constraints. This may be explained from the biggest relative difference between mean AT and FI scores for this configuration from Section 3.1. However, adding SemMedDB training instances (with minimum frequencies 2 and 3) seems to lead to a potentially more desirable compromise with recall around 96% and precision over 80%. The results for count based CMF models were disappointing as shown in Table 3. There is no clear pattern as to how the scores are spread with regards to different configurations and overall performance is all across the board inferior when compared to BMF models especially with substantially lower precision values.

### Table 2: Performance results of BMF models for approved indications in repoDB

<table>
<thead>
<tr>
<th>Unallowed cases</th>
<th>Threshold</th>
<th>P</th>
<th>R</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>0.00001</td>
<td>0.812</td>
<td>0.942</td>
<td>0.8727</td>
</tr>
<tr>
<td>50%</td>
<td>0.00001</td>
<td>0.811</td>
<td>0.942</td>
<td>0.8723</td>
</tr>
<tr>
<td>75%</td>
<td>0.00001</td>
<td>0.812</td>
<td>0.941</td>
<td>0.8721</td>
</tr>
<tr>
<td>100%</td>
<td>0.005</td>
<td>0.723</td>
<td>0.981</td>
<td>0.8328</td>
</tr>
</tbody>
</table>

### Table 3: Performance results for CMF models over repoDB

<table>
<thead>
<tr>
<th>Unallowed cases</th>
<th>Threshold</th>
<th>P</th>
<th>R</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>0.12162</td>
<td>0.714</td>
<td>0.957</td>
<td>0.8185</td>
</tr>
<tr>
<td>50%</td>
<td>0.08999</td>
<td>0.698</td>
<td>0.968</td>
<td>0.8119</td>
</tr>
<tr>
<td>75%</td>
<td>0.00009</td>
<td>0.692</td>
<td>0.973</td>
<td>0.8093</td>
</tr>
<tr>
<td>100%</td>
<td>0.005</td>
<td>0.703</td>
<td>0.963</td>
<td>0.8131</td>
</tr>
</tbody>
</table>

4 DISCUSSION

As CDR efforts continue to rise, it is critical to have benchmarking datasets that are realistic in terms of representation of both approved indications and failed indications. repoDB is first of its kind dataset that creates such an opportunity to conduct comparative evaluations of CDR methods on a publicly available gold standard dataset.

4.1. Main takeaways. Matrix completion through NMF based low-rank approximation is an effective method for CDR based solely on datasets of previously approved drugs and corresponding indications. Actually, in this manuscript, we only use public data sources of treatment relations in the form of hand curated UMLS Metathesaurus relations and those extracted with NLP from PubMed citations (from SemMedDB). As such, these are imperfect resources (especially SemMedDB) and may not necessarily constitute FDA approved drugs. Results still show that among recoverable ATs from repoDB, we achieved F-scores close to 90% with the highest F-score achieved with just UMLS relations as input. Using both SemMedDB and UMLS relations helps achieve a better compromise between precision and recall with over 96% recall at 80% precision. The mean NMF score for FIs is at least twice as large as that for RIs, indicating that FIs are indeed much tougher to distinguish from ATs compared with randomly generated pairs. A critical enabler was the encoding of unallowed pairs (derived with incompatible semantic type constraints from UMLS semantic network) as zeros imposing additional structural constraints on the input matrix to be approximated. However, imposing constraints from all unallowed pairs could be detrimental by leading to a 3% recall gain with a 10% precision drop. Experiments showed that introducing 25% of the zeroes from unallowed pairs leads to better outcomes and is computationally less expensive. Count based models that consider frequency from SemMedDB substantially underperform compared with simpler binary models. Overall, NMF based methods applied to carefully curated external knowledge
sources constitute a practical approach towards CDR.

Next we discuss some examples of correct predictions made by our approach. In our training dataset we see the drug vincristine treating malignant neoplasms, follicular lymphoma, and Hodgkin disease and another drug doxorubicin treating the general condition of malignant neoplasms. After matrix completion, we saw high values of 0.89 and 0.93 for the entries (doxorubicin, follicular lymphoma) and (doxorubicin, Hodgkin disease) respectively, which are approved indications in repoDB that were never encountered in training data and were blank cells before the training process. Similar new correct predictions are also made for (bleomycin, follicular lymphoma) and (bleomycin, Hodgkin disease). Next, although, the count based CMF method underperformed overall, there were cases where it lead to correct predications when the binary approach did not. For example, (betamethasone, berylliosis) and (bleomycin, malignant head and neck neoplasm) are approved indications that were missed by the BMF approach but are recovered by the CMF method. Thus there may be some complementary traits in how the BMF and CMF approaches predict that need further examination toward building an ensemble method.

We set out to explore reasons for errors — false positives (FPs) and false negatives (FNs) — incurred by the NMF models in the context of information available about the corresponding drug–disease pairs. To this end, we examined connectedness of FP and FN pairs in the SemMedDB graph, which essentially conveys the potential shared context between associated entities. In our prior work, we identified graph patterns over the SemMedDB graph that are highly indicative of treatment relations using model coefficients of a logistic regression (LR) model. For FPs of the NMF model, we noticed that there were tens of thousands of highly predictive short paths (length ≤ 3) connecting the corresponding drug and disease CUIs indicating that there are many shared neighbors; some of this neighborhood information is encoded in the input matrix, which could have led to positive predictions. This is also not surprising given FPs are essentially failed cases in repoDB but were deemed plausible enough for researchers to launch clinical trials. For FNs, we found relatively fewer and sometimes no such predictive paths in SemMedDB connecting associated entities. For example, for the approved repoDB indication (Dexamethasone, Branch retinal vein occlusion with macular edema), the drug and disease were not connected in the SemMedDB graph using LR model’s top predictive patterns. Without much shared context, NMF appears to struggle to elicit positive indications for such pairs. We plan to pursue a more detailed error analysis involving physician experts, which may yield additional insights on potential reasons for errors.

4.2. Limitations and future work. This current effort is not without a few limitations, which also point to interesting future research directions for CDR experiments with repoDB.

- The method in this paper is clearly not a silver bullet for CDR. repoDB does enable excellent benchmarking but in general scientists are often looking at a particular disease that they want to treat. Hence, for disease specific CDR, more sophisticated methods involving gene expression datasets and methods that consider integration of various modalities of information specific to the disease may be needed, as indicated in other prior efforts (e.g., Nagaraj et al. for cancer). However, our method can be an effective initial step in pruning the space of candidates before more sophisticated methods that require more complex modeling and disease specific information can be applied.

- As discussed earlier, the count based CMF models’ performance was underwhelming (Table 3) when compared with the binary models even though the counts capture additional information about prior knowledge being incorporated into the input matrix. One reason for this could be that we simply employed raw frequencies of treatment predications in SemMedDB instead of standardizing counts using well-known methods (e.g., mean centering, min-max scaling, log transformation). Using raw frequencies may have lead to potential ill-conditioning that needs to be countered with appropriate pre-processing and/or using more sophisticated methods. These experiments will be part of future extensions of our work.

- Matrix completion methods cannot fill a row that does not have at least one nonzero entry. In our case, this means, a drug for which we do not have at least one known treatment relation cannot be linked to new indications with NMF. However, this can be remedied by moving from matrices to tensors with additional relations between entities connected with other predicates including prevents, diagnoses, affects, and causes. Using tensor factorization, even

---

5The LR model, while being effective, is computational prohibitive at times given the explosion of numbers of paths connecting entities in a large graph such as that built from SemMedDB. Our foray into NMF is motivated by these efficiency constraints of the graph pattern based approach.
for a drug with no existing treatment relations, using multi-hop indirect connections, it is possible to elicit a new indication. Similarly, with recent deep learning advances, embedding nodes and edges of the larger SemMedDB graph (including edges arising from other predicates besides treats) with graph neural networks can offer a different way for knowledge base completion. We intend to pursue these directions in the immediate future.

4.3. Benchmarking. To enable future comparisons with our results, we provide the validation/test set splits of repoDB drug–disease pairs used in this study: https://github.com/bionlproc/nmf-repoDB-benchmarking. This will be important for direct comparisons by other researchers using the repoDB dataset, especially given we had to resort to using a subset of repoDB (owing to issues with lack of training instances for certain drugs without a single human vetted treatment relation).

5 CONCLUSION

With valuable time and cost savings in the offing, CDR efforts are expected to increase in the future. With lack of datasets modeling both positive and failed indications, it is encouraging to notice that datasets such as repoDB are being created. However, it is also important to start comparing methods against such datasets for robust assessments of different methods. In this paper, matrix completion through NMF was used to directly predict repoDB approved indications by using publicly available treatment relations. F-scores close to 90% were obtained with various training configurations with this method showing its strong potential for practical applications. Validation and test splits of repoDB used as part of this effort are made available to facilitate direct comparisons with our results by other researchers in the CDR community. More sophisticated methods such as tensor factorizations and neural graph embeddings may hold the promise of recovering novel indications for drug compounds that have not yet been approved for any known conditions. We believe this is the first attempt to employ repoDB for CDR purposes and hope that this will trigger more attempts to pursue this line of work toward rigorous benchmarking.

Acknowledgements

We are grateful for the support of the U.S. National Library of Medicine through NIH grant R21LM012274 and also thankful for partial support offered by the U.S. National Center for Advancing Translational Sciences via grant UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We also acknowledge the Ministry of National Education, Republic of Turkey, for providing financial support to Gokhan Bakal with full scholarship for his doctoral studies. HK was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

[8] Trevor Cohen, Dominic Widdows, Clifford Stephan, Ralph Zinner, Jeri Kim, Thomas Rindflesch, and Peter Davies. Predicting high-throughput screening results with scalable literature-based discovery methods. CPT:
pharmacometrics & systems pharmacology, 3(10):1–9, 2014.


Feasibility Assessment of a Pre-Hospital Automated Sensing Clinical Documentation System

Sean M. Bloos, Candace D. McNaughton, MD PhD MPH, Joseph R. Coco, MS, Laurie L. Novak, PhD MHSA, Julie A. Adams, PhD, Robert E. Bodenheimer, PhD MS, Jesse M. Ehrenfeld, MD MPH, Jamison R. Heard, Richard A. Paris, Christopher L. Simpson, MA, Deirdre M. Scully, and Daniel Fabbri, PhD

1 Department of Biomedical Informatics, 2 Department of Emergency Medicine, 3 Department of Anesthesiology
Vanderbilt University Medical Center, Nashville, TN
4 Department of Electrical Engineering & Computer Science
Vanderbilt University, Nashville, TN
5 School of Electrical Engineering & Computer Science
Oregon State University, Corvallis, OR

Abstract
Clinical documentation in the pre-hospital setting is challenged by limited resources and fast-paced, high-acuity. Military and civilian medics are responsible for performing procedures and treatments to stabilize the patient, while transporting the injured to a trauma facility. Upon arrival, medics typically give a verbal report from memory or informal source of documentation such as a glove or piece of tape. The development of an automated documentation system would increase the accuracy and amount of information that is relayed to the receiving physicians. This paper discusses the 12-week deployment of an Automated Sensing Clinical Documentation (ASCD) system among the Nashville Fire Department EMS paramedics. The paper examines the data collection methods, operational challenges, and perceptions surrounding real-life deployment of the system. Our preliminary results suggest that the ASCD system is feasible for use in the pre-hospital setting, and it revealed several barriers and their solutions.

Introduction
Military and civilian medics are responsible for retrieving, stabilizing, and transporting the wounded and ill to a trauma facility. Accurately documenting medical care during transport is complicated for several reasons, including limited staff in the vehicle and ongoing care requirements for trauma and critically ill medical patients. Because it is not possible to manually document all activity during transport as it occurs, medics often first give a brief verbal report to the receiving facility staff, followed by documentation completed after the hand-off has occurred. The initial verbal report typically includes chief complaint, mechanism of injury, vital signs, and procedures performed. This verbal report may be supported by brief notes written on the patient, a scrap piece of paper, the medic’s glove, or in some cases relying only on the medic’s memory. While accurate transmission of this information is essential for safe care transitions and plays a key role in the receiving team’s treatment plans, reporting may be incomplete and include inaccuracies. As a result, communication, safety, and patient care may be impacted.

Communicating some types of patient information can be done successfully with a verbal report alone (e.g., chief complaint, mechanism of injury, age, gender). However, specifics regarding the sequence of procedures performed, medication dosage and timing, and specific vital sign ranges are difficult to accurately recall from memory given the high-intensity setting of trauma care. This information is essential for optimal care management, resource allocation, and triage planning.

An Automated Sensing for Clinical Documentation (ASCD) system leverages a combination of sensors to passively collect data, from which algorithms are used to create an abbreviated care record that describes the procedures performed. This record is designed to be generated in real time, or near real time, and transmitted to providers as a supplement to the verbal handoff. The goal of this system is to increase the accuracy and detail of clinical information transmitted to clinical providers and teams, particularly in high-acuity and trauma settings, without requiring the provider to actively produce the documentation. An ASCD system can be used in a range of environments from civilian patient transport, military patient transport, and operating rooms.
This paper reports on a feasibility study of an ASCD system for patient transport in a civilian metropolitan Emergency Medical Service (EMS). Specifically, this paper outlines the equipment used, the configuration of the equipment in a civilian ambulance, perception of medics wearing devices, data collection processes, and interfaces with the trauma facility. The system was deployed with the Nashville Fire Department (NFD) EMS in partnership with Vanderbilt University Medical Center, a level I trauma center in Nashville, Tennessee, which receives a high volume of acute trauma patients. Cameras were not deployed in the system during the pilot because of concerns regarding patient and provider privacy, although their use would be expected to increase system accuracy.

**Background**

High quality healthcare requires effective and accurate communication among providers, especially during transitions across care settings. The dynamic nature of care by medics in the pre-hospital setting can make it difficult to document procedures in real time and communicate vital clinical information to hospital providers. In their handoffs to emergency department (ED) staff, paramedics dedicate 75% of verbal reports to patient demographics and presenting signs/symptoms and ~7% to pre-hospital treatments, even though pre-hospital treatments and clinical course largely drive resource allocation and treatments upon arrival. Various approaches have been considered to automatically document clinical care including speech-to-text and in-person/virtual scribes. However, given environmental noise, limited vehicle space, and the need for the medic to focus on patient care rather than documentation, alternative approaches to automated documentation systems are desired. Moreover, given the number of and fluctuation in transport vehicles and number of trips, ease of device acquisition and installation is important.

Off-the-shelf sensors offer an opportunity to gather the data necessary to produce documentation, while also being readily available and cost-effective. These sensors include electromyography (EMG) sensors, inertial measurement units (IMUs), such as accelerometers, and cameras. EMG sensors collect data that measure electrical currents during muscular contraction, which can be used to identify different neuromuscular activity. IMUs combine data from accelerometers, gyroscopes, and magnetometers to report the orientation and angular rate of a body. Cameras can track the medic and patient in the scene. These data feeds can be fed into machine learning algorithms for analysis and documentation generation.

For this pilot study, the NFD assisted in the evaluation of the system. NFD provides fire protection and emergency medical care for 533 square miles and transports patients to numerous hospitals within the metropolitan Nashville area. In 2018, NFD responded to approximately 130,000 calls. Of note, NFD EMS protocols for airway management do not include the use of medication for rapid sequence intubation.

**Methods**

An ASCD system for patient transport in a pre-hospital setting has many components. First, a series of sensors are deployed in a clinical environment, in this case EMG sensors and accelerometers measuring provider movement. Second, these data feeds are aggregated and analyzed by machine learning systems to detect the clinical activity that are performed. Third, from the detected activity, an abbreviated care record is produced that can be transmitted to upstream care providers. This paper describes the methods around the feasibility of deploying sensors on clinicians to collect data for analysis.

A range of sensors were considered for the feasibility test ranging from medic worn sensors, mounted cameras, to radio-frequency identification chips on each device and/or medication. Given the desire of a rapid, cost-effective, and simple deployment that protected patient privacy, sensors worn by medics were selected. Sensors worn by medics that were utilized in this study included Myo armbands, which collect EMG data and inertial measurement data (e.g., accelerometers), and Apple Watches, which collect yaw, pitch, roll, and acceleration data.

After obtaining institutional review board approval from the Vanderbilt University Human Subjects Protection Program, physicians, nurses, and paramedics were recruited to perform over 45 hours of procedures in Vanderbilt’s Center for Experimental Learning and Assessment (CELA), which is a high-fidelity simulation center. Each participant was consented and then asked to perform designated procedures multiple times over the duration of a three-hour period using a simulation patient. Data collected from the simulations were used to refine the sensors, data collection systems, and the algorithm prior to real-world deployment. These simulations included use of cameras.
The ASCD system was then deployed in partnership with NFD over a 12-week period in 2019, without cameras. Written informed consent was obtained from the participating paramedics. Paramedic shifts were selected based on the availability of paramedics and research staff. At the beginning of each shift, a trained researcher who is also a paramedic equipped one paramedic with sensors, specifically two Apple Watches and two Myo armbands. In order to collect and transmit the collected information, a laptop and two cell phones were also carried by the research observer (see Data Collection). The data were aggregated into a single server for future analysis.

During the NFD paramedic’s 12-hour shift, the trained researcher observed and recorded all clinical activity with a custom time-motion capture system the “start” and “stop” times for targeted procedures performed inside the ambulance (Table 1). These procedures were chosen based on focus groups with EMS personnel and common procedures performed in the ambulance. Automated time-motion capture was not feasible given the subjectivity and complexity of the procedures. Procedures performed outside of the ambulance were excluded from the recorded observations due to the distance from the laptop and Bluetooth receivers to the Myo armbands and iPhones. The research observer did not participate in patient care.

For the subset of patients transported to Vanderbilt University Medical Center (VUMC) adult emergency department (ED), observations also included the handoff between the team of paramedics and the ED team. Procedures and interventions performed during the ED visits were also documented. At the conclusion of each shift, the paramedic who wore the technology completed a debrief survey. These surveys featured a user-centered design approach, considered the context of use, specific requirements, and areas of design optimization. Between each observation, the ASCD equipment was cleaned using SaniWipes® and all components of the system were charged. No substantial damage was received to the equipment during use.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Start</th>
<th>Stop</th>
<th># Times procedure observed during deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administer Intramuscular (IM) Medication</td>
<td>Medication taken out of the box</td>
<td>Needle withdrawn from patient</td>
<td>0</td>
</tr>
<tr>
<td>Administer Intraosseous (IO) Medication</td>
<td>Medication taken out of the box</td>
<td>Finished flushing with saline</td>
<td>0</td>
</tr>
<tr>
<td>Administer Intravenous (IV) Medication</td>
<td>Medication taken out of the box</td>
<td>Finished flushing with saline</td>
<td>3</td>
</tr>
<tr>
<td>Apply Pressure to Stop Bleeding</td>
<td>First applying pressure to artery</td>
<td>Hand leaves wound</td>
<td>0</td>
</tr>
<tr>
<td>Bagging</td>
<td>Touches bag</td>
<td>Hand leaves bag</td>
<td>0</td>
</tr>
<tr>
<td>Blood Pressure Measurement</td>
<td>Touches blood pressure cuff</td>
<td>Hands leaves blood pressure cuff</td>
<td>13</td>
</tr>
<tr>
<td>Chest-Tube Insertion</td>
<td>Touches scalpel</td>
<td>Apply tape/bandage to secure chest tube</td>
<td>0</td>
</tr>
<tr>
<td>Chest-Tube Preparation</td>
<td>Has hemostat in one hand and tube in the other</td>
<td>Hands leave hemostat</td>
<td>0</td>
</tr>
<tr>
<td>Chest-Tube Suturing</td>
<td>Touches needle-driver</td>
<td>Apply tape/bandage</td>
<td>0</td>
</tr>
<tr>
<td>Combat Gauze</td>
<td>Touches gauze</td>
<td>Patient's wrapped extremity is put down</td>
<td>0</td>
</tr>
<tr>
<td>Combat Tourniquet</td>
<td>Touches tourniquet</td>
<td>Writes time on tourniquet</td>
<td>0</td>
</tr>
<tr>
<td>CPR (Respiratory Support)</td>
<td>Touches patient's head</td>
<td>Hands leaves patient's head</td>
<td>0</td>
</tr>
<tr>
<td>CPR (Compressions)</td>
<td>Starts first compression</td>
<td>Hands leaves patient</td>
<td>0</td>
</tr>
<tr>
<td>Draw Medication</td>
<td>Removes syringe from packaging</td>
<td>Syringe leaves medication vial</td>
<td>1</td>
</tr>
<tr>
<td>ECG Lead Application</td>
<td>Touches ECG electrodes</td>
<td>Hands leave patient after final electrode applied</td>
<td>6</td>
</tr>
<tr>
<td>Intubation with Endotracheal (ET) Tube</td>
<td>Touches laryngoscope</td>
<td>Removes syringe after inflating the cuff of ET tube</td>
<td>0</td>
</tr>
<tr>
<td>Intraosseous (IO) Access</td>
<td>Touches IO drill</td>
<td>IO is secured to the patient</td>
<td>0</td>
</tr>
<tr>
<td>Intravenous (IV) Access</td>
<td>Touches IV from package</td>
<td>Taped IV down</td>
<td>11</td>
</tr>
</tbody>
</table>
Supraglottic Airway (SGA) | Touches SGA | Removes syringe after inflating the cuff | 0
---|---|---|---
Oropharyngeal Airway (OPA) | Touches OPA | Hand leaves OPA after placement | 0
Pulse Oximetry Monitoring | Touches pulse oximetry monitor | Hands leave pulse oximetry monitor | 1
Splinting | Touches splint | Wrapped extremity is put down after splint is applied | 0
Swab Area with Alcohol | Touches alcohol prep | Alcohol prep is out of user’s hand | 6
Swap Vial with Alcohol | Alcohol towelette is removed from packaging | Alcohol towelette is removed from vial | 0
Tie Tourniquet (preparation for IV access) | Touches IV tourniquet | Hands leaves IV tourniquet | 5
Vital Monitoring | Touches stethoscope or thermometer | Hands leave stethoscope or thermometer | 5
Wound Dressing | Touches gauze package | Patient’s wrapped extremity is put down | 0

Data Collection
Transport of Equipment: The following equipment was housed in a Pelican case during transport: 2 Myo Armbands, 2 Apple Watches (Series 3), 2 Apple iPhone 7s, AT&T Unite Express 2 WiFi hotspot, 2013 Apple MacBook Air, and their accompanying chargers (Figure 1). The equipment was delivered at the beginning of each shift and kept with a member of the research team.

**Figure 1**: Pelican case setup

Electromyography: Thalmic Labs Myo gesture control armbands were used as the source for EMG data collection. Myo armbands were connected to the MacBook Air using a Bluetooth connection. After the armbands were pulled onto the forearms, they were calibrated by having the paramedics position their hands as shown in Figure 2.

**Figure 2**: Myo neutral position (A) to Myo sync position (B) to calibrate the armbands
**Accelerometer:** One Apple Watch Series 3 was worn on each wrist with the watch-face outward and was paired with an Apple iPhone 7. Both iPhones were stored in a secure cabinet in the rear of the ambulance. The iPhones were connected to a virtual private network (VPN), which was necessary to securely send data to a VUMC server. We originally used the PulseSecure application to access the VPN, but then the system migrated, causing the switch to F5 Access. To establish an internet connection, a AT&T Unite Express 2 WiFi hotspot was used.

**Data Flow (Figure 3):** Both the left and right Myo Armbands were connected to the laptop using a Bluetooth connection. The laptop was connected to the mobile hotspot over a VPN. The left and right Apple Watches were paired with a corresponding iPhone 7, which also used the hotspot WiFi.

![Figure 3: Flow of data inside the ambulance](image)

**Data Transfer:** At the conclusion of each shift, the Myo data collection files were transferred from the laptop to a shared cloud account (VUMC Box) for review by the data analysts. Due to the large file sizes, the transfer was completed using VUMC WiFi instead of the hotspot.

**Notification of VUMC research team:** An application, Life360, was installed onto one of the iPhones. A geo-fence was set up around VUMC, which notified the clinical researcher at the VUMC ED when the phone was detected within 800 feet of the ambulance bay. In addition, the research observer would send a message via Life360 to the clinical researcher when enroute to VUMC. The clinical researcher met paramedics and the observer in the ED to document the handoff to hospital staff.

**Documentation of Targeted Procedures:** The research team thoroughly discussed the method for documenting the targeted procedures prior to deployment of the ASCD system. The research observer needed a quick and accurate system that allowed for the research observer to note any discrepancies in how the paramedics actually started and stopped each procedure. Variations in procedural equipment or protocols could cause the start and stop times to differ. Original discussions included methods such as (i) keeping field notes and manually keeping track of time, or (ii) creating an iPhone application that allowed the observer to press a start/stop button with pre-selected procedures. The most significant challenge with the former was accuracy of procedure times. For the latter, it was determined that an application with pre-selected options did not allow the observer much versatility to add other comments and raised concerns about accidental selection of a start or stop time. Ultimately, a simple Python logging application was used that allowed for free-text entries and recorded time stamps upon entry.
Research documentation began at the point in which the patient was loaded into the ambulance. During each transport, the observer typically sat in the captain’s or “airway” seat (rear facing), located behind the head of the patient, and documented from that position. Once the patient was loaded into the ambulance, the research observer opened the laptop, ensured proper connectivity of the system, and began recording procedures as they happened in real time. For example, if the paramedic was going to start an IV, the observer would type “IV start” into the log application when the paramedic touched the IV start kit. At the end of the procedure, indicated by taping down of the IV, the observer would type “IV end”. These were used as the gold standard for “start” and “stop” times for each procedure of interest (Table 1; Table 2 for representative data records).

<table>
<thead>
<tr>
<th>Procedure Occurred</th>
<th>Procedure Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-01-09 12:42:13.030240</td>
<td>12 lead start</td>
</tr>
<tr>
<td>2019-01-09 12:42:38.984363</td>
<td>12 lead end</td>
</tr>
<tr>
<td>2019-01-09 12:43:07.886611</td>
<td>correction 3 lead</td>
</tr>
<tr>
<td>2019-01-09 12:44:53.242214</td>
<td>albuterol tx start</td>
</tr>
<tr>
<td>2019-01-09 12:45:11.890404</td>
<td>mask applied</td>
</tr>
<tr>
<td>2019-01-09 12:45:59.294573</td>
<td>tourniquet</td>
</tr>
<tr>
<td>2019-01-09 12:46:02.672012</td>
<td>IV start</td>
</tr>
<tr>
<td>2019-01-09 12:46:27.348737</td>
<td>IV in</td>
</tr>
<tr>
<td>2019-01-09 12:49:31.370297</td>
<td>IV procedure END</td>
</tr>
<tr>
<td>2019-01-09 12:49:54.562330</td>
<td>IV attempt fail</td>
</tr>
<tr>
<td>2019-01-09 12:57:29.503253</td>
<td>12 lead</td>
</tr>
<tr>
<td>2019-01-09 12:57:48.126144</td>
<td>12 lead end</td>
</tr>
<tr>
<td>2019-01-09 12:58:58.214071</td>
<td>albuterol tx end</td>
</tr>
<tr>
<td>2019-01-09 12:59:40.566805</td>
<td>check lung sounds</td>
</tr>
</tbody>
</table>

Paramedic Debriefs
After each observed shift, the paramedic completed a debrief survey. The research observer also provided feedback to the research team regarding lessons learned in the field, barriers encountered, and feedback obtained. Survey responses from the paramedic participants were entered into REDCap, a secure web application designed for creating and managing online surveys and databases15.

Establishing Rapport with the Paramedic Participants
The research observer was a paramedic, and this allowed him to build a productive rapport with the paramedic team. The observer arrived at the beginning of each shift, either 0530 or 1730 hours for a day or night shift, respectively. The observer typically stayed for 6 hours, depending on call volume. During busier shifts, the observer stayed with the crew for an extended period of time to collect more data.

Results
Over seven observations, two paramedics wore the system for a total of 49 hours. We observed the transport of 16 patients to 6 different hospitals. Information after handoff for six patients was obtained (Table 3). Using the first procedure logged as a start time and the end of the last procedure logged as an end time, the median time of active treatment during transport was 8 minutes and 15 seconds (standard deviation of 5 minutes and 24 seconds).

<table>
<thead>
<tr>
<th>Patient #</th>
<th>ESI Score</th>
<th>ED Disposition</th>
<th>Chief Complaint / Mechanism of Injury</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>Transfer to specialty care</td>
<td>Suspected ingestion</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Transfer to specialty care</td>
<td>Overdose</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Discharge</td>
<td>Auto vs. pedestrian</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Unknown</td>
<td>Generalized weakness</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>Transfer to specialty care</td>
<td>Suicidal ideations</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>Discharge</td>
<td>Intoxication/Chest Pain</td>
</tr>
</tbody>
</table>

Table 4 describes the different barriers encountered during the data collection process and the approaches taken to mitigate the issues. The most common challenge was ensuring that all components of the system were properly
functioning and collecting data simultaneously. Specific solutions are listed below. Additionally, the placement of the laptop in the ambulance and ASCD sensors on the medic are depicted in Figure 4.

### Table 4: Challenges encountered during data collection and their solutions

<table>
<thead>
<tr>
<th>Barriers</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermittent interruption in Myo Armband connectivity to laptop</td>
<td>Laptop location moved to the head of the stretcher, under the patient’s head (Figure 4)</td>
</tr>
<tr>
<td>Script programs stopped recording when the laptop lid was shut</td>
<td>Installed disable lid sleep widget</td>
</tr>
<tr>
<td>Intermittent interruption in Apple Watch data collection</td>
<td>Implemented a live feedback system to visualize interruptions in data collection</td>
</tr>
<tr>
<td>Insufficient hotspot data for data collection</td>
<td>Data use was monitored proactively via web portal; ~1GB was needed per 6-hour observation</td>
</tr>
<tr>
<td>Confusing Apple Watch start/stop application</td>
<td>Start/stop feature changed from a tapping mechanism to a slide bar</td>
</tr>
<tr>
<td>Concerns of marrying pre-hospital observations to correct paramedic-to-ED handoff</td>
<td>Process developed to ensure consecutive subject data entry; relative time (since the start of paramedic shift) used to identify patients</td>
</tr>
<tr>
<td>Myo armbands intermittently vibrate if they are unsynced (caused by displacement of armband)</td>
<td>Paramedics were cautioned that this may occur, and they attempted to not desync the armbands; this vibrating functionality will be removed in future trials.</td>
</tr>
<tr>
<td>Systemwide VPN upgrade for VUMC users</td>
<td>We were forced to switch the VPN connection on the laptop; it had no apparent effect on data collection</td>
</tr>
<tr>
<td>Original hotspot data plan was canceled by the carrier for an unknown reason</td>
<td>Observations were delayed until we were able to obtain a new hotspot and data plan</td>
</tr>
</tbody>
</table>

**Current Documentation Techniques by Paramedic Participants**

During field observations, the majority of documentation occurred in the ambulance. For “non-critical” patients (e.g., stable vital signs), paramedics typically used the charting software installed on their Toughbook® laptops to document items such as past medical history, current medications, drug allergies, and demographic information. In addition, the cardiac monitors used by NFD had the ability to store vital signs such as blood pressure, heart rate, oxygen saturation, and respiration rate. This log could then be uploaded directly to the patient care report following the transport. During the care of patients who required more attention, or were more critical, documentation typically took place in the form of the paramedic writing on the glove of their non-dominant hand. In other situations, paramedics documented some procedures from memory, which is standard care practice for EMS providers. Patient care documentation into the EMS formal charting system typically occurred post-trip arrival. Paramedics indicated that charting for each patient took approximately 30 minutes in total, which is longer than the transport time but in line with other work assessing time that clinicians spend documenting.16.

**Figure 4**: Laptop in ambulance (A) and a medic wearing two Myo armbands and two Apple Watches (B)
Operational Challenges
Although the study was designed around most patients being transported to VUMC, patients were transported to six different hospitals. This occurred for multiple reasons, including EMS diversion (hospital not accepting patients arriving by EMS), patient request to be transported to a specific hospital, proximity, and the clinical condition of the patient. Additionally, several scheduled observation periods were canceled due to mechanical malfunction of the ambulance, personnel illness, and work schedule changes.

Communication of Procedures to Receiving Facility Staff
Six handoffs between paramedics and ED clinical teams at VUMC were observed. Two handoffs occurred at ED triage, in which paramedics gave a verbal report to a triage nurse. Verbal reports from the paramedic to the ED staff largely consisted of the chief complaint and signs and symptoms. During the handoff of a level II trauma patient, the paramedic relayed all information regarding procedures (IV placement, medication administration, vital signs, and cervical collar application) to the trauma team, which was composed of physicians, nurses, and other ED staff.

General Feedback from Paramedics (Table 5)
The paramedic participants indicated that they could wear the armbands for the duration of a 12-hour shift with no anticipated difficulties. Participants reported that the armbands were tight but did not restrict their overall movement or interfere with patient care. An impression on the paramedic’s skin of the armbands was observed for ~30 minutes following the removal of the armbands. Participants suggested considering other devices that might be more comfortable or fit in clothing. Other comments included that this technology would be useful in situations where there are critical patients or multiple patients.

Table 5: Results from paramedic questionnaires

<table>
<thead>
<tr>
<th>Factor</th>
<th>Response (N=7 Field Observations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability to wear entire shift</td>
<td>Yes – 7/7</td>
</tr>
<tr>
<td>Perceived comfortableness</td>
<td>Neutral – 1/7</td>
</tr>
<tr>
<td></td>
<td>Slightly Uncomfortable – 6/7</td>
</tr>
<tr>
<td>Likelihood to wear entire shift</td>
<td>Unlikely – 1/5</td>
</tr>
<tr>
<td></td>
<td>Likely – 2/5</td>
</tr>
<tr>
<td></td>
<td>Extremely Likely – 2/5</td>
</tr>
<tr>
<td>Issues with devices interfering with uniform</td>
<td>“They do not interfere with clothing”</td>
</tr>
<tr>
<td></td>
<td>“No issues”</td>
</tr>
<tr>
<td></td>
<td>“There are no complications with uniform and armbands”</td>
</tr>
<tr>
<td>Feelings regarding devices tracking movements</td>
<td>“I do not have any concerns”</td>
</tr>
<tr>
<td></td>
<td>“I do not see any real problems with the devices tracking my movements”</td>
</tr>
<tr>
<td>Overall experience</td>
<td>“They get more uncomfortable the longer they are on. Even with all of the links taken out, they are a bit tight”</td>
</tr>
<tr>
<td></td>
<td>“It has been a good experience”</td>
</tr>
<tr>
<td>Perceived feasibility of automated documentation</td>
<td>“I think it’s completely feasible to have it automatically document time on action to improve documentation accuracy”</td>
</tr>
<tr>
<td></td>
<td>“I feel like it would be very helpful in the pre-hospital setting with exact times and interventions”</td>
</tr>
<tr>
<td>Perceived usefulness of automated documentation</td>
<td>“It would be helpful on calls that require more hands on the patient where you don’t have time to document as you go”</td>
</tr>
<tr>
<td></td>
<td>“It would be useful when we are dealing with a critical patient or have multiple patients on the same scene. It would also be helpful to have this information to help give a report to the ED”</td>
</tr>
</tbody>
</table>

Discussion
To our knowledge, there have been no attempts to create and deploy an automated clinical documentation system using this range of sensors. We identified challenges surrounding logistics, connectivity, evaluation, and perception by the paramedics. The results of this pilot study support the feasibility of using an Automated Sensing for Clinical Documentation System in the pre-hospital setting. The paramedic debriefs suggest that the equipment used in this study can be deployed in a large study with more paramedics, allowing for more data to be captured.

The system’s aim is to support and supplement the documentation and upstream communication processes that occur during and after transport. For example, the ASCD system would be particularly useful in settings where verbal
communication is limited, such as casualty evacuation from the battlefield or multiple simultaneous patient arrivals. Creation of an abbreviated care record with timestamps identifying which procedures were performed can enable upstream hospital providers to more effectively provide care. Specifically, the precision of such timestamps can aid in decisions regarding subsequent medication administration and other judgments surrounding patient viability.

The pilot highlighted several barriers to the system’s practical deployment. Most of the barriers encountered were logistical in nature. For example, ensuring that the data recorded continuously throughout the duration of the observation was of the most significant, which we addressed through various engineering enhancements. Modifications to the design of the system were made to make it more user-friendly, such as indicating when the system was actively recording data. These steps facilitated more accurate and complete data collections.

We anticipate that deployment of the system into the field would result in greater variation than the simulated procedures recorded in CELA, and therefore, introduce subsequent error into our model. Detection of such error will allow us to refine the current algorithm used for the identification of targeted procedures.

Methods for Future Data Analysis
This paper specifically does not discuss the algorithms used to convert the collected sensor data to an abbreviated care record. Briefly, the algorithms take the IMU and EMG data as input to generate summary statistics (e.g., entropy, power, etc.) and analyze temporal patterns. These data are then fed into classifiers to predict which procedure is being performed, if any. Some data cleaning are also performed such as removing the gravity vector and removing vehicle vibration.

Future processing will include employing a deep neural network using both convolutional and recurrent layers with memory trained classifiers in hopes for a greater efficacy. We plan to move towards a real-time analysis using discrete, rolling windows for classification such that the start and stop of the event is not known by the classifier, which will be the case in real-world deployments.

Limitations
Our study intentionally limited the number of paramedic participants (2) in order to optimize the system and to identify and address system failures. Having a single observer and paramedic participant during the initial deployment allowed us to identify and address challenges before expanding the study. The second limitation can be attributed to the unknown nature of pre-hospital care. Because of the unpredictable nature of pre-hospital care, there were some procedures for which we were not able to collect any data.

Conclusion
This paper reported on the lessons learned from deploying an automated sensing clinical documentation system in a real-world environment. It discussed challenges of configuring equipment, collecting data, and dealing with failures. Many incremental steps were taken to reach the goal of a working system that could be safely deployed in the field and collect data, without interfering with care and patient privacy. Future work will analyze how well the algorithms are able to correctly identify which procedures were performed during transport.
References


One-Way and Round-Trip Analysis Demonstrates Surprising Limitations of Standards-Based Terminology Maps

Steven H. Brown MS MD¹², Loren Stevenson MD¹, Daniel J. Territo MS¹, John Kilbourne MD¹, Jonathan R. Nebeker MD¹³, Holly Miller RN MS¹, Michael J. Lincoln MD¹³

¹Department of Veterans Affairs, ²Vanderbilt University Medical Center, Nashville Tennessee, ³University of Utah, Salt Lake City Utah

Abstract

The informatics community has a long-standing vision of freely flowing and highly re-usable patient-specific clinical data that improves care quality and safety. We sought to evaluate the extent to which a standards-based mapping approach is sufficient to support semantic interoperability. We simulated large-scale clinical data transmission and measured semantic success between VA and DoD systems via one-way testing (OWT) and round-trip testing (RTT). Simulations were accomplished via SQL queries and production standards-based maps for medications, allergies, document titles, vitals and payers. Success rates for mapping local codes to national standards varied from 62.5% for DoD document titles and medications, to 100% for VA and DoD vital signs. Successful, one-way testing was considerably lower, ranging from 8.52% to 62.7%. Round-trip success rates were lower still, ranging from 1.7% to 76.3%. We present an error framework, lessons learned, and proposed mitigating steps to enhance standards-based semantic interoperability.

Introduction

Calls for the meaningful sharing of standards-based electronic health data are more than 25 years old.¹⁻³ These reports, and others more recent, share a common vision of freely flowing and highly re-usable patient-specific clinical data that improves care quality and safety. The HITECH act of 2009 mandated and achieved widespread national adoption of electronic health records (EHRs) by hospitals and providers.⁴⁻⁵ HITECH-act-driven EHR adoption in tandem with associated regulations requiring the use of specific data standards has helped the US take significant steps towards improving healthcare via shared data.⁶⁻⁷

Demands for health data interoperability between the Department of Defense (DoD) and the Department of Veterans Affairs (VA) share a similar timeline with more challenging technical and organizational requirements.⁸⁻¹⁰ For example, The National Defense Authorization Act (NDAA) for Fiscal Year 2014 required the Secretaries of the Department of Defense and the Department of Veterans Affairs to ensure that “Not later than October 1, 2014 all health care data contained in the Department of Defense AHLTA and the Department of Veterans Affairs VistA systems shall be computable in real time and comply with the existing national data standards.”¹¹ The 2014 NDAA further required “transition to modern, open architecture frameworks that use computable data mapped to national standards to make data available for determining medical trends and for enhanced clinician decision support.”

The Department of Veterans Affairs and the Department of Defense responded to this legislation by developing and deploying the Joint Legacy Viewer (JLV).¹²,¹³ JLV is a product that shares and displays patient data at the point of care. JLV extracts data from legacy EHRs (e.g., VistA and AHLTA) via Corporate Data Warehouses (CDW) and applies data maps between internal data representations and designated national standards to facilitate exchange. DoD and VA also assigned the joint Interagency Clinical Informatics Board (ICIB) to define “all health care data” and necessary standards to meet the NDAA mandate. The ICIB subsequently identified 23 data “domains” (e.g., medications, immunizations) and relevant standards for their exchange. Each Department developed and maintains its data maps independently. JLV is deployed at all VA Medical Centers and Military Treatment facilities. Between March and November of 2018, users logged in to JLV approximately 300,000 times and viewed approximately 1,200,000 records per month.

Saitwal and Hussain both note that data mapping is a more complex activity than might be assumed at first blush.¹⁴,¹⁵ Achieving the long-standing vision of meaningful sharing of standards-based electronic health data
(i.e., semantic interoperability) requires the successful marshaling, mapping, transmission, re-mapping and use of transferred information. This complex pathway requires solutions for mapping and marshaling challenges (e.g., as noted by L’Amore et al). We undertook this study to assess impacts on semantic interoperability of applying standards-based clinical data maps to a large, real-world clinical data set.

**Methods**

**Selected Domains**

We selected the five domains that the JLV is capable of integrating standards-based data for: Document titles (LOINC), medications (RxNorm), allergens (RxNorm + SNOMED CT), vital signs (LOINC) and payers (ASC X12).

**Data Preparation**

We developed a “test region” in VA’s Corporate Data Warehouse for this evaluation. The CDW test region copied the data structures and de-identified clinical data from the CDW production system servicing JLV. Data was de-identified by authorized CDW staff using approved standard “test region” encryption methods prior to being made available to the evaluation team. The same approved encryption methods were applied to the DoD Corporate Data Repository (CDR), ensuring that patients could be matched across data sets but not identified. We calculated counts of all data instances and unique codes for each domain.

**Mapping-Success Analysis**

Mapping to standards is a necessary prerequisite for standards-based interchange. We calculated the numbers and percentages of local codes that were mapped to a standards-based mediating code. We also calculated the number of mediating codes mapped to sets of local codes (i.e., one mediating code mapped to many local codes) stratified by set size.

**Data Transmission Simulation**

We simulated data transmission between VA and DoD systems using two distinct patterns: one-way testing (OWT) and round-trip testing (RTT). One-way testing separately simulated the transmission of VA data to DoD and the transmission of DoD data to VA. In each case, the data was transformed through each map set one time (i.e., the sending system’s local codes were mapped to standards-based “mediating” codes and then into the receiving system’s local codes). Round-trip testing simulated bi-directional data transmission (e.g., from VA to DoD and back again). Simulations were accomplished via SQL queries performed against test region tables using JLV production maps.

![Figure 1](image_url) Simulated Data Flows. VA-DoD Data One-Way Testing (solid arrows, steps 1-3): 1) VA data element retrieved from test region and matched to VA-Standards map (one-to-many). 2) Standards-based data element(s) sent to Standards-DoD Map. DoD map output compared to DoD data elements in test region (one-to-many). Round-trip testing return (dashed arrows, steps 4-6) reverses the process.
**Scoring**

Algorithmically exact matches between input and output data were scored as fully successful without additional human review. One-way testing success occurred when both the sending and receiving agencies shared the same mediating code (which have mappings to local codes and surface forms), regardless of the receiving system’s local code or surface form. If the mediating code appeared more than once in the receiving systems mapping tables (tied to a different local code), each match was counted as a success. This choice was made because sending and receiving local code systems could not be expected to have the same surface forms or internal identifiers.

Round-trip testing success occurred when the local data sent was received back in precisely its original form—i.e., matching codes and surface forms. An unsuccessful RTT means that the precise, original meaning was not sent back to the originating partner. Full success RTT rate is calculated with the numerator equals the number of full successes and the denominator equals “intent to interoperate” (size of domain). We also documented the total number RTT sent messages (“tries”) created by the many-to-one mappings.

**Results**

**Data Extraction**

We extracted de-identified patient-instance data from the VA Corporate Data Warehouse and the DoD Clinical Data Repository. Table 1 shows the number of clinical data elements and unique codes extracted for each domain for each Department.

**Table 1. Data Set Overview**

<table>
<thead>
<tr>
<th></th>
<th>VA Data Instances</th>
<th>VA Unique Codes</th>
<th>DoD Data Instances</th>
<th>DoD Unique Codes</th>
<th>Total Data Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payers</td>
<td>235,569</td>
<td>50</td>
<td>3,126,159</td>
<td>29</td>
<td>3,361,728</td>
</tr>
<tr>
<td>Vitals</td>
<td>252,299,085</td>
<td>17</td>
<td>8,331,186</td>
<td>11</td>
<td>260,630,271</td>
</tr>
<tr>
<td>Docs</td>
<td>290,334,004</td>
<td>3,438</td>
<td>1,046,252</td>
<td>80</td>
<td>291,380,256</td>
</tr>
<tr>
<td>Meds</td>
<td>62,932,809</td>
<td>16,338</td>
<td>28,387,162</td>
<td>66,769</td>
<td>91,319,971</td>
</tr>
<tr>
<td>Allergens</td>
<td>715,073</td>
<td>3,225</td>
<td>267,237</td>
<td>4,346</td>
<td>982,310</td>
</tr>
</tbody>
</table>

**Mapping to Standards**

The percentage of mapping successes (i.e., mapping a local code to a standards-based mediating code) for each Department’s clinical data is detailed in Table 3 and Table 4. They varied from 62.5% for DoD document titles to 100% for VA and DoD vital signs. Seven of ten mapping-success proportions exceeded 90%.

We found that in many instances multiple local codes were mapped to a single, standards-based mediating code. For example, two VA source elements “CEFADROXIL 1GM TAB” and “CEFADROXIL MONOHYDRATE 1GM TAB” were mapped to a single RxNorm semantic clinical drug “Cefadroxil 1000 MG Oral Tablet”. Table 2 shows the counts of instances where sets of local codes mapped to a single mediating code. DoD medications, DoD allergens and VA document titles had more than 40 bins (e.g., long tails). There were no cases of multiple standards-based mediating codes mapping to a single local code in any domain.
One-Way Testing Results

Table 3 shows VA→DoD one-way testing results, and Table 4 shows DoD→VA one-way testing results. In Table 3, VA to DoD one-way testing success as percent of total unique document titles, vital signs, medications and allergens were 8.5%, 29.4%, 46.8%, and 26.5% respectively. Payers were successful in no instances.

Table 3. VA→DoD One-Way Test Results for Unique Identifiers

<table>
<thead>
<tr>
<th>Domain</th>
<th>VA Unique Codes</th>
<th>VA-SDO Mapped</th>
<th>VA-SDO Mapped %</th>
<th>Success N</th>
<th>Success % of total</th>
<th>Success % of mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>3,438</td>
<td>3,416</td>
<td>99.36%</td>
<td>293</td>
<td>8.52%</td>
<td>8.58%</td>
</tr>
<tr>
<td>Payers</td>
<td>50</td>
<td>50</td>
<td>100.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>17</td>
<td>17</td>
<td>100.00%</td>
<td>5</td>
<td>29.41%</td>
<td>29.41%</td>
</tr>
<tr>
<td>Medications</td>
<td>16,338</td>
<td>12,204</td>
<td>74.70%</td>
<td>7,648</td>
<td>46.81%</td>
<td>62.67%</td>
</tr>
<tr>
<td>Allergens</td>
<td>3,225</td>
<td>2,905</td>
<td>90.08%</td>
<td>856</td>
<td>26.54%</td>
<td>29.47%</td>
</tr>
</tbody>
</table>

In Table 4, showing DoD→VA one-way testing results, DoD to VA one-way testing success as percent of total unique document titles, vital signs, medications and allergens were 25%, 45.5%, 20.6%, and 42% respectively. Payers, again, were successful in no instances.

Table 4. DoD→VA One-Way Test Results for Unique Identifiers

<table>
<thead>
<tr>
<th>Domain</th>
<th>DoD Codes</th>
<th>DoD-SDO Mapped</th>
<th>DoD-SDO Mapped %</th>
<th>Success N</th>
<th>Success % of total</th>
<th>Success % of mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>80</td>
<td>50</td>
<td>62.50%</td>
<td>20</td>
<td>25.00%</td>
<td>40.00%</td>
</tr>
<tr>
<td>Payers</td>
<td>29</td>
<td>28</td>
<td>96.55%</td>
<td>0</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>11</td>
<td>11</td>
<td>100.00%</td>
<td>5</td>
<td>45.45%</td>
<td>45.45%</td>
</tr>
<tr>
<td>Medications</td>
<td>66,769</td>
<td>41,746</td>
<td>62.52%</td>
<td>13,757</td>
<td>20.60%</td>
<td>32.95%</td>
</tr>
<tr>
<td>Allergens</td>
<td>4,346</td>
<td>3,946</td>
<td>90.80%</td>
<td>1,825</td>
<td>41.99%</td>
<td>46.25%</td>
</tr>
</tbody>
</table>
**Mediating Code Impact**

A mediating code is an SDO code that VA and DoD each map to for any given concept, enabling a one-way trip. Table 5 represents the actual impact of the SDO mediating codes on one-way testing success. This table describes actual instances of mediation, given the number of shared concepts (“mediating codes”). For example, VA and DoD had a partially disjointed, overlapping set of vital signs (e.g., VA audiometric “vitals” not shared by DoD). However, the Departments shared five vital signs (e.g., heart rate). Those five shared vitals mediated 65.06% of VA and 56.30% of eight DoD measurements. A total of 164 million VA vital signs and 8.3 million DoD vital sign instances were examined. Across all five domains, considering both directions, mediation rates (excluding payors) ranged between 26.4% and 91.8%.

**Table 5. Mediating Code Impact on Actual Instances of Clinical Data**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Mediating Codes</th>
<th>VA Data Instances</th>
<th>VA Mediated Instances</th>
<th>% VA Mediated</th>
<th>DoD Data Instances</th>
<th>DoD Mediated Instances</th>
<th>% DoD Mediated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>14</td>
<td>290,334,004</td>
<td>86,925,863</td>
<td>29.94%</td>
<td>1,046,252</td>
<td>932,716</td>
<td>89.15%</td>
</tr>
<tr>
<td>Payers</td>
<td>0</td>
<td>235,569</td>
<td>0</td>
<td>0.00%</td>
<td>3,126,159</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>5</td>
<td>252,299,085</td>
<td>164,155,351</td>
<td>65.06%</td>
<td>8,331,186</td>
<td>4,690,531</td>
<td>56.30%</td>
</tr>
<tr>
<td>Medications</td>
<td>5,731</td>
<td>62,932,809</td>
<td>57,778,448</td>
<td>91.81%</td>
<td>28,387,162</td>
<td>23,052,269</td>
<td>81.21%</td>
</tr>
<tr>
<td>Allergens</td>
<td>789</td>
<td>715,073</td>
<td>451,343</td>
<td>63.12%</td>
<td>267,237</td>
<td>70,480</td>
<td>26.37%</td>
</tr>
</tbody>
</table>

**Documents**: 14 LOINC codes mediated 29.9% of all VA document titles and 89.2% of all DoD documents titles in the data set when adjusted for frequency.

**Payers**: Payers have had a mapping file mismatch resulting in 0 mediations. Each Department used different versions of the ANSI X12 standard that were not interoperable (VA used payer typology; DoD used ANSI ASC X120-like codes plus “additions” that were unique to their system).

**Vital Signs**: Five LOINC mediated 65.1% of all VA vital signs and 56.3% of all DoD vital signs in the data set.

**Medications**: 5,731 RxNorm codes mediated 91.8% of VA and 81.2% of DoD medications in the dataset.

**Allergens**: 789 RxNorm and SCT codes mediated 63.1% of VA allergens and 26.4% of DoD allergens in the data set.

**Round-Trip Testing Results**

In round-trip testing, the actual results, excluding Payors, varied between 1.7% and 76.3%. But even when standards were the same (e.g., RxNorm for Medications), differences in formularies (value sets used by each Department) and non-exact mappings meant that round-trip testing could still fail. The table illustrates the results for all five domains and origins from either VA (VA→DoD→VA) or DoD (DoD→VA→DoD). For example, Medication round trips succeeded no more than 29% of the time (in that case, the direction was VA→DoD→VA).

**Table 6. Round-Trip Testing Results for Unique Identifiers**

<table>
<thead>
<tr>
<th></th>
<th>VA→DoD→VA Tries</th>
<th>VA→DoD→VA Success %</th>
<th>DoD→VA→DoD Tries</th>
<th>DoD→VA→DoD Success %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payers</td>
<td>0</td>
<td>0.00%</td>
<td>29</td>
<td>0.00%</td>
</tr>
<tr>
<td>Vitals</td>
<td>5</td>
<td>29.41%</td>
<td>11</td>
<td>45.45%</td>
</tr>
<tr>
<td>Docs</td>
<td>500</td>
<td>1.69%</td>
<td>974</td>
<td>51.33%</td>
</tr>
<tr>
<td>Meds</td>
<td>22,771</td>
<td>29.42%</td>
<td>259,959</td>
<td>8.76%</td>
</tr>
<tr>
<td>Allergens</td>
<td>2,120</td>
<td>76.26%</td>
<td>11,072</td>
<td>19.15%</td>
</tr>
</tbody>
</table>
Error Framework

Certain types of errors repeatedly occurred, and so we established a corresponding error framework with the following categories:

**Plausible Mismap:** Codes for same or equivalent concepts mapped to different yet arguably plausibly correct targets.
Example: VA maps to “Body Weight”, DoD maps to “Body Weight, Measured”

**Mapping Error:** Mapping errors can be pernicious but perhaps are not inevitable. Errors may occur at the time of map creation or emerge as the result of terminology and formulary changes.
Example: VA “HEPATOLOGY C&P EXAMINATION CONSULT” incorrectly mapped to LOINC 3\textcircled{3}9867-6 instead of LOINC 3\textcircled{3}8967-6. Error surveillance, including versioning and configuration management, can mitigate this risk. Importantly, each Department or exchange partner should consider using a common methodology for mapping. For example, they might develop and share a common “style guide” for mappers. Consistent use of tooling that helps prompt and even enforce consistent mapping can be useful. Finally, an Independent Verification and Validation (IV&V) of both partners’ mappings can be performed, ideally by the same group and using the same IV&V methodology.

**Granularity Error:** The granularity of a term is a measure of its specificity and refinement. A coarsely granular concept may encompass many finely granular terms (e.g., subclasses). Map granularity can be a design choice based on intended use or reflect patient-instance data-knowledge limits (e.g., patient x is allergic to penicillin(s) (class) vs. Benzathine Penicillin G (precise ingredient)). Example: VA maps document titles to 3416 LOINC terms, DoD maps to 50.

**Formulary Error:** Sending and receiving systems may not utilize or be able to represent the same clinical “items” (e.g., medications, lab tests). Final step expressivity deficits may limit transmission success (i.e., nothing to map to). This could be addressed by maintaining common set of items or by having methods to determine the useful similarity of items. For example, drug-allergy interactions could be computed using an RxNorm semantic clinical drug component rather than the precise RxNorm semantic clinical drug. Example: DoD has “PEAK FLOW” as a vital sign; VA does not.

**Information Model Error:** Differences in information model may make single term mapping impossible. Example: VA maps to BP, DoD maps to systolic BP and diastolic BP. Both systems represent blood pressure metadata such as cardiac cycle, patient position, cuff size, and artery sampled entirely differently.

**Many-to-One Explosions:** Many-to-one mapping explosions induced a large number of errors. The number of RTT tries observed corresponds precisely to the product of the number of codes at each step of the transmission (Table 7). The observed RTT success percentage corresponds to the reciprocal of the number of local codes in the many to one relationship (e.g., 3 codes sent $\times 33\%$ success on return). For RTT tries this is an $O(n^3)$ process. Semantic interoperability cannot be achieved until viable solutions to this issue are implemented.

<table>
<thead>
<tr>
<th>VA NDF Sent</th>
<th>RxNorm Sent</th>
<th>DoD Local Pivot</th>
<th>RxNorm Return</th>
<th>VA NDF Return</th>
<th>Total RTT Tries</th>
<th>RTT Success</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>Cefadroxil 1000 MG Oral Tablet</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>Cyclosporine, modified 100 MG Oral Capsule [Gengraf]</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>Zolpidem tartrate 10 MG Oral Tablet</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>Pentaerythritol Tetrานitrate 10 MG Oral Tablet</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>18</td>
<td>6</td>
<td>Pilocarpine Hydrochloride 40 MG/ML Ophthalmic Solution</td>
</tr>
</tbody>
</table>
We undertook qualitative analysis of many VA NDF codes to one RxNorm Code instances. One consistent cause of many-to-one mappings relates to “Quantity Factors” for liquid preps. A “Quantity Factor” is the amount of medicine packaged in a vial or syringe. For example, VA “DOXORUBICIN HCL 200MG/VIL (PF) INJ” and VA “DOXORUBICIN HCL 10MG/VIL (RDF) INJ” both mapped to RxNorm “Doxorubicin Hydrochloride 2 MG/ML Injectable Solution”. The version of RxNorm used in this study was not able to represent and distinguish quantity factors. New versions of RxNorm now have this capability.

We performed limited quantitative analysis of many-to-one mappings in the domain of medications. Medications provide an extreme example, Medication RTT from VA \( \text{\rightarrow} \) DoD \( \text{\rightarrow} \) VA yielded 54,632 failures: 8,690 from unrecognized medications, and 45,942 from recognized many-to-one mappings. Table 7 demonstrates that many-to-one mappings impact round trips \( O(n^3) \) more than one-way success rates \( O(n^2) \).

**Discussion**

A high percentage of mappings between clinical data and national standards existed within each Department, yet one-way and round-trip testing success percentages based on mediating codes between Departments were much lower. This occurred in part because the Departments share partly disjoint value sets (e.g., no DoD counterpart for VA audiometric “vitals”). The problem is further exacerbated because the Department mappings were done independently and not entirely consistently: there was no examination between Departments when apparently very similar terms, and likely identical concepts, were mapped to different targets. These two factors alone meant that one-way trips would not always equal the proportion of overall mappings. Finally, the combinatorial explosion of one-to-many mappings made successful trips, particularly round trips, even worse. Therefore, for example, 99% of unique document titles in VA were mapped to national standards but the 14 mediating LOINC codes only mapped 8.6% of VA’s unique identifiers in one-way trips to DoD. Likewise, for vital signs, 100% of DoD’s and VA’s local identifiers were mapped to national standards but the 5 mediating LOINC codes matched only 29.4% of VA vital signs and 45.5% of DoD vital signs in one-way tests. Frequency-adjusted mediation proportions were higher because the most commonly occurring concepts were more likely to be shared between the value sets, and because the independent Department experts were more likely to have mapped similarly for such common items. In addition to the error framework presented in the results section, we learned several other lessons about potential challenges in map-based interoperability, namely:

**Standards “Suitability”:** The similarity of the Standard’s information model to the deployed system information model affects success rate. In cases where the deployed information model includes more data than the Standards information model, one-to-many mappings can result, causing decreased full-trip success rate even when all data is known or knowable.

**Style:** Standards may have more than one way to reasonably represent the same concept. This can result in plausible mismappings that must be managed by sending or receiving systems.

**Intent:** Maps, like terminologies(16), should be created with a specific purpose in mind. Maps created for one purpose (e.g., many drugs to one category for drug allergy interactions) may not work well for other purposes (e.g., granular pharmacy inventory management).

**Change Management:** The quality of maps can degrade over time. Two types of change must be managed: terminology versions and change in value sets (formularies) to be mapped and exchanged. We have found, in maintaining JLV maps, that we can expect ca. 12-15% annual change in active domains such as medications. This results from standards-bodies’ changes (e.g., retired RxNorm CUIs), formulary changes and new drug entities.

As a result of lessons learned, we propose several approaches to improve semantic interoperability. First, whenever possible, standards should be used natively in EHRs. In our experience, mapping is a recurrently expensive and inherently lossy process. Reducing the number of mapping translations reduces interoperability complexity and cost. Barriers to native use of standards, potentially including expressivity and update frequency, must be better understood. Second, ongoing standards quality improvement, for example, the addition of the “quantity factor” to RxNorm, can improve interoperability. Additional standards extensions and improvements should be use-case
driven and could contribute reductions in many-to-one, granularity, and formulary errors. Third, we recommend that EHRs should deploy entire terminologies and maintain separate purpose-specific subsets (e.g., ordering medications). The whole terminology could underpin other uses (e.g., CDS and reporting) and provide context for inbound, standards-based data. Additional benefits may accrue by using aggregation hierarchies and class definitions rather than enumerated lists to drive CDS, reporting, and other uses. Chu et al. have found this approach to reduce complexity and knowledge engineering time while increasing completeness coverage of clinical phenotypes.(17) We postulate that this approach may prove to be more resilient than enumerated lists for interoperating with inbound, standards-based data. Finally, our fourth suggestion is to develop robust mapping tools, common mapping procedures (at least a shared “desiderata” of mapping) and undertake strong independent validation and verification (IV&V) to best improve future results.

Conclusion

As our present results show, even when mapping is funded, prioritized and reliably completed, significant clinical information can be lost in one-way and round-trip transmission. Understanding this fact and addressing the underlying causes are important for continued progress in semantic interoperability and achieving the opportunities for improving healthcare quality and safety that it offers.

Clinicians increasingly expect computer-driven decision support, intelligent prompts, error detection and truly comparable data. They want to sum the lifetime Adriamycin dose across a continuum of care, even when the legacy systems in that continuum do not use common drug formularies, much less units of measure. Healthcare systems and practices that acquire systems should better understand the limitations that relate to data quality and interoperability and press for ongoing improvement. We suggest that system and standards developers focus on several problems: 1) the expense, inaccuracy and probable unsustainability of purely mapping-based approaches; 2) the need to support, maintain and natively use in HIT, common and high quality terminology standards; 3) migrating to common information models for complex clinical expressions and conclusions; 4) the idea that HIT software is evanescent and the clinical data are the gold nuggets that matter to patients and clinicians. Interoperability will not happen without ongoing investment and quality improvement cycles. We hope this paper has brought to light new opportunities for progress.

References


8. United States. General Accounting Office, GAO. Federal health care: increased information system sharing could improve service, reduce costs: briefing report to the Chairman, Committee on Veterans Affairs, House of Representatives. 1993.


Determination of Marital Status of Patients from Structured and Unstructured Electronic Healthcare Data

Brian T. Bucher, MD, MS1, Jianlin Shi, MD, MS1, Robert John Pettit, BS1, Jeffrey Ferraro, PhD2, Wendy W. Chapman, PhD1, Adi Gundlapalli, MD, PhD1,3
1 University of Utah School of Medicine, Salt Lake City, UT, USA; 2 Intermountain Healthcare, Salt Lake City, UT, USA; 3 VA Salt Lake City Healthcare System, Salt Lake City, UT, USA

Abstract
Social Determinants of Health, including marital status, are becoming increasingly identified as key drivers of health care utilization. This paper describes a robust method to determine the marital status of patients using structured and unstructured electronic healthcare data from a single academic institution in the United States. We developed and validated a natural language processing pipeline (NLP) for the ascertainment of marital status from clinical notes and compared the performance against two baseline methods: a machine learning n-gram model, and structured data obtained from the electronic health record. Overall, our NLP engine had excellent performance on both document-level (F1 0.97) and patient-level (F1 0.95) classification. The NLP Engine had superior performance compared with a baseline machine learning n-gram model. We also observed a good correlation between the marital status obtained from our NLP engine and the baseline structured electronic healthcare data (κ 0.6).

Introduction
Social Determinants of Health (SDoH) are set of constructs that are becoming increasingly identified as key moderators of health-related outcomes1. For example, socioeconomic status, housing availability/stability, and social support affect the health of patients and should not be overlooked when investigating access to health care utilization and outcomes. SDoH can affect healthcare via a number of direct mechanisms including behavioral and disease risk factors, access to and processes of care, and quality of health care received. Recently, the United States Centers for Medicare and Medicaid Services (CMS) has proposed financial penalties to hospitals with high rates of 30-day readmission compared to the national averages2. Although this program’s goal is to improve the quality of health care provided, it fails to take into account SDoH that may drive differences in readmission, thus penalizing hospitals who preferentially care for high-risk patient populations.

Social support and social relationships are an important type of SDoH that can aid patients in access to healthcare and support during acute and chronic health care episodes. In particular, marital status is a traditional construct of social support and has been shown to have a positive effect on healthcare outcomes. Being married or in a domestic relationship has been shown to improve outcomes after hip replacement, heart transplantation, and even positively impact mortality3-5. With regards to resource utilization, being married has been shown to be associated with shorter hospital length of stay and hospital readmission6,7. Being widowed or lacking a partner has also been associated with increased risk of hospital readmission8. It is important to note that marital status is not only dichotomous (married vs not married), and may also include being separated, divorced, or widowed. Furthermore, in contemporary times, domestic partners are also considered under this status.

The challenge in studying the effect of SDoH on health-related outcomes is the lack of reliable data available to clinicians and researchers. SDoHs are inconsistently collected in the electronic health record (EHR) and vary between institutions9. Common SDoH that are routinely collected include: age, gender, and race/ethnicity. However marital status, while commonly collected in structured form as part of an intake to a healthcare facility, has the potential to change over time and may not be updated appropriately or in a timely manner to reflect the changes. In addition, given the constantly changing definitions of marriage and domestic partnership, the structured field may be inadequate to capture a patient’s own view of marital status.

Free text clinical notes provide a rich source of SDoH that has the potential to be accessed using natural language processing techniques (NLP)10. The advantage of clinical notes includes rich text features which provide context around SDoH. In addition, patients may disclose marital status more freely to providers and health care workers who document the social history in clinical notes. In these situations, the advantage of clinical notes is the ability to identify changes in marital status over time.
Previous work by our group and others have demonstrated the value of NLP in the identification of SDoH from free-text clinical notes. In a study from the VA Healthcare System, we developed an NLP system to extract mentions of “social support,” “housing situation,” and “living alone”\textsuperscript{10}. The performance of the system on a held-out test set showed an F1 score of 0.90 for social support, 0.61 for housing situation, and 0.81 for living alone. In a separate study, Navathe et al. develop a rules-based NLP system from 500 annotated clinical notes from a single academic institution to identify the following SDoH: tobacco use, alcohol use, drug abuse, depression, housing instability, fall risk, and poor social support\textsuperscript{11}. They demonstrated excellent performance with F1-scores ranging from 0.75 for Alcohol Use to 0.94 for Poor Social Support. In addition, they demonstrated the value of NLP-identified terms in predicting readmission risk compared to structured data alone.

To our knowledge and literature review, no previous studies have utilized NLP for the ascertainment of marital status. This is likely due to the belief that marital status is commonly collected in structured form in the electronic health record. However, as marital status may change over time, the accuracy of the structured data collection is unknown as it can be prone to error if it is not updated appropriately. The goal of the current study is to describe the development and validation of a rules-based NLP pipeline for determination of marital status from clinical notes. Our rules-based pipeline is then compared to a machine learning n-gram model and marital status obtained from structured data fields in the electronic health record. We subsequently describe the advantages and disadvantages of ascertainment of marital status from both structured data and NLP.

**Methods**

**Study Design**

We performed a retrospective study of patients treated at the University of Utah Health Sciences Center from 2015-2017. The University of Utah has maintained an electronic health record capable of comprehensive electronic data capture since 2013 and the data are stored in an enterprise data warehouse (EDW). Given we are interested in marital status determination, we limited our corpus to clinical Social Work Notes as these notes were highly likely to contain references to marital status.

**Structured Data Acquisition**

In the University of Utah EDW, marital status is treated as a patient level structured data field. This data field is updated on demand with updates in the electronic health record. Therefore marital status is not linked to encounter based information. To work around this limitation we obtained the log records for marital status changes to determine the date of marital status changes at the point of encounters.

**Reference Standard Corpus Creation**

Because there is only one structured field in the EHR for marital status and the structured data field is not linked to each encounter, we consider clinical notes to be a better representation of a patient’s true marital status for each encounter. We obtained 23,794 Social Work notes from 4716 patients between 2015-2017. We performed manual document-level annotation on 865 notes to create a reference standard corpus\textsuperscript{12}. We trained two annotators who performed document level annotation for marital status. The document level categories were: Single, Domestic Partner, Married, Separated, Divorced, Widowed. Overall, the inter-annotator agreement was 0.96. Any discrepancies were adjudicated between the two annotators to create a reference standard corpus. We then split the corpus into a development set (378 notes) to develop our rules-based pipeline or train the machine-learning model, and a blind validation set (487 notes) to measure performance.

**Knowledge Base Development**

To design our knowledge base for our domain, we started by creating a roadmap of the terms and relationships that define the underlying concept of marital status using real-world clinical text (annotations) from the development corpus as a guide. In addition, we enriched the ontology with terms using the Unified Medical Language System\textsuperscript{13}. The knowledge base was revised iteratively based on expert opinion and chart review and was used in our rules-based NLP pipeline.

**NLP Engine Development**

We utilized EasyCIE, a lightweight, rules-based NLP tool, which supports quick and easy implementation of clinical information extractions\textsuperscript{14}. EasyCIE uses a set of highly optimized and customizable NLP components built above n-Trie (a fast rule processing engine)\textsuperscript{15}, including sentence segmentation\textsuperscript{16}, named entity recognition\textsuperscript{17}, and context detection\textsuperscript{18}, feature inferencer, and document inferencer (Table 1).
Table 1. NLP components and corresponding functionality of EasyCIE.

<table>
<thead>
<tr>
<th>NLP Components</th>
<th>Functionality Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section Detector</td>
<td>Identify the sections, e.g., History of Present Illness, Family History</td>
</tr>
<tr>
<td>Sentence Segementer</td>
<td>Detect sentence boundaries</td>
</tr>
<tr>
<td>Named Entity Recognizer</td>
<td>Identify target concepts using dictionaries</td>
</tr>
<tr>
<td>Context Detector</td>
<td>Attach the context information as feature values to the corresponding target concepts</td>
</tr>
<tr>
<td>Attribute Inferencer</td>
<td>Create mention-level conclusions based on target concepts and corresponding attributes</td>
</tr>
<tr>
<td>Document Inferencer</td>
<td>Create a document-level conclusion from the corresponding mention-level conclusions</td>
</tr>
</tbody>
</table>

EasyCIE leverages the knowledge base built above with information extraction models (IEM), including a term mapping IEM: a semantic representation of target concepts (“married”, “divorced”, “wife,” “husband”, etc.) and the corresponding contextual modifiers. The context IEMs for this study include four types: 1) Negation, whether a target concept is negated or affirmed, e.g. “not married”; 2) Certainty, whether a target concept certainly or uncertainly indicates the marital status, e.g. “marital status unknown”; 3) Temporality, whether a target concept indicates present, historical, or future, e.g. ”wife recently died”; 4) Experiencer, whether a target concept is referring to a patient or a patient’s family member, e.g. “met with patient’s daughter and her husband.” An annotation schema was defined from these IEMs. For example, “patient has been married for 5 years” would be encoded as Target Concept: Married, Negation: affirmed, Certainty: certain, Temporality: current, Experiencer: patient. Based on the values for these annotations in the document, EasyCIE applies rules to infer the final marital status for each document: Single, Domestic Partner, Married, Separated, Divorced, Widowed, Unknown. The rules were developed from the development corpus by manual review of reference standard annotations and iterative error analyses. The source code for EasyCIE can be found at https://github.com/jianlins/EasyCIE_GUI. The knowledge base and configuration file for ascertainment of marital status can be found at https://github.com/jianlins/EasyCIE_Hub.

Machine Learning Baseline Development

Given the lack of previous studies using NLP to determine marital status from clinical notes, we sought to compare the performance of the rules-based NLP system to a baseline machine learning. We developed an ngram-classifier trained using unigram and bigrams (1-2 word windows as word features) from the development set, excluding stop words. Our final feature set included 1488 binary n-gram features. We then created a word vector for each document in the development and validation set. We used the manual annotations as the reference standard to develop and validate the machine learning approach. We trained a Random Forest Classifier on the development set and used 10-fold cross-validation to tune the regularization parameters. Once we had optimal performance on the development set, we tested the performance of the machine learning classifier on the blind validation set.

Analysis

Performance of both the machine learning classifier and our NLP engine was measured by precision, recall, and F1 score for classifying documents compared against the reference standard annotations. The definitions are shown below

\[
\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}
\]

\[
\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}
\]

\[
F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]

We compared the performance of our NLP approach to the baseline machine learning approach using McNemar’s Test defined below where b and c are the discordant positive and negative classified documents between the NLP and machine learning classifier and \(\chi^2\) is the test statistic:

\[
\chi^2 = \frac{(b - c)^2}{b + c}
\]
We compared the performance of our NLP engine against the structured data fields using Cohen’s Kappa defined below where \( p_o \) is the observed agreement and \( p_e \) is the expected agreement by chance:

\[
\kappa = \frac{p_o - p_e}{1 - p_e}
\]

**Results**

**Patient Cohort and Annotation**

We obtained structured marital status from 4716 patients treated from 2015-2017. Based on the structured EHR data, the majority of patients in the cohort were married (55%), followed by single (23%), divorced (9%), widowed (7%), Domestic Partner (2%), and Separated (2%). Overall, marital status was unknown in 2% of the cohort. Since marital status can change over time, we examined the number of times the marital status changed in the structured data fields (Table 2). Overall, 16% of patients had at least one change in marital status with 2% of patients changing four times or more. Of the changes, 581 (78%) occurred on the same day, indicating a number of changes may be related to duplicate entries or updates to existing marital status.

**Rules-based NLP Engine**

Given the changes and possible uncertainty in the structured data fields, we sought to develop and validate an NLP engine to identify a patient’s marital status from clinical Social Work notes and compare the performance of our NLP engine to the structured data fields.

**Table 2.** Changes in Marital Status in the Structured EHR data fields.

<table>
<thead>
<tr>
<th>Number of Changes</th>
<th>Count (Percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3951 (84%)</td>
</tr>
<tr>
<td>1</td>
<td>247 (5%)</td>
</tr>
<tr>
<td>2</td>
<td>242 (5%)</td>
</tr>
<tr>
<td>3</td>
<td>194 (4%)</td>
</tr>
<tr>
<td>4+</td>
<td>82 (2%)</td>
</tr>
</tbody>
</table>

The performance of our NLP engine and machine learning approach on the validation set is shown in Table 3. Overall the NLP engine had significantly better performance compared to the baseline machine learning approach (F1-Score 0.97 vs 0.63, p-value <0.001). Examining individual classes, the rules-based approach had excellent performance (F1 score>0.90 in all individual class with “Separated” having the best performance. Compared to the baseline machine learning approach, the rules-based NLP approach did not have significantly better performance for “Divorced” or “Separated” classes. This is likely related to the low number of patients who met these criteria.

We then compared the performance of the NLP engine to that of the structured data field for patient-level marital status classification on the validation set by aggregating across patients and eliminating unknowns. (Table 4) Compared to the NLP engine, the structured data field showed decreased classification accuracy on patient level classification; however, this difference was not statistically significant (p=0.4).

**Table 3.** Baseline Machine Learning and Rules-Based NLP Engine Performance

<table>
<thead>
<tr>
<th>Class</th>
<th>Machine Learning Baseline</th>
<th>rules based NLP Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.75</td>
<td>0.92</td>
</tr>
<tr>
<td>Single</td>
<td>0.90</td>
<td>0.4</td>
</tr>
<tr>
<td>Domestic Partner</td>
<td>0.6</td>
<td>0.21</td>
</tr>
<tr>
<td>Married</td>
<td>0.75</td>
<td>0.82</td>
</tr>
<tr>
<td>Separated</td>
<td>1.0</td>
<td>0.25</td>
</tr>
<tr>
<td>Divorced</td>
<td>0.88</td>
<td>0.85</td>
</tr>
<tr>
<td>Widowed</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Overall</td>
<td>0.84</td>
<td>0.56</td>
</tr>
</tbody>
</table>

\(a\). Micro-Averaged F1 Score
Table 4. Patient-Level Classification Performance on the Validation Set

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rules Based NLP</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>0.74</td>
<td>0.43</td>
<td>0.51</td>
</tr>
<tr>
<td>Structured</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Error Analysis

We analyzed the failures of our NLP engine in the validation set and characterized the errors in Table 5. Overall the majority of errors (42%) were related to lack of appropriate context assignment due to mentions of partners/spouses of relatives. The next most common errors occurred due to lack of negation (25%) or failure of document level inferencing due to multiple conflicting mentions of marital status.

Comparison of NLP vs Structured Data Field

Having established the excellent performance of our NLP Engine, we processed the entire corpus of 23,784 social work notes with the NLP engine and compared classification by NLP against the structured data fields. Since marital status may change over time, we aggregated notes by Year and Quarter to determine the NLP-derived marital status for that particular time. The NLP and structured labels are shown in Table 6. Overall the NLP-engine was unable to determine the marital status on 28% of patients, compared to only 3% for the structured data (p<0.0001).

Table 5. Natural Language Processing System Error Analysis

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Count (Percent)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context Inference</td>
<td>5 (42%)</td>
<td>“Patient lives with his sister Sarah and her husband”</td>
</tr>
<tr>
<td>Document-Level Inference</td>
<td>3 (25%)</td>
<td>“Patient is divorced but lives with his girlfriend”</td>
</tr>
<tr>
<td>Negation</td>
<td>3 (25%)</td>
<td>“Patient was never married”</td>
</tr>
<tr>
<td>Named Entity Recognition</td>
<td>1 (8%)</td>
<td>“Marital Status: W”</td>
</tr>
</tbody>
</table>

We next compared the correlation of NLP-derived labels with those in the structured data. Overall the Cohen’s κ between the NLP-derived variable was 0.37. Excluding the unknown labels, the Cohen’s κ improved to 0.60. The pairwise agreement between the NLP-derived variable and the structured variable is shown in Figure 1. Overall there was a good agreement for Married (73%), and modest agreement for Domestic Partner (45%), Divorced (46%) and Widowed (51%). The largest disagreement was for Single. Of the patient labeled as Single by the structured data fields but not by NLP, NLP assigned the following marital status: domestic partner (14%), Married (10%), Divorced (9%), Separated (1%) and Widowed (1%).

Table 6. The Distribution of NLP and Structured Data Labels in Full Corpus

<table>
<thead>
<tr>
<th>Class</th>
<th>NLP</th>
<th>Structured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>787 (10%)</td>
<td>2361 (30%)</td>
</tr>
<tr>
<td>Domestic Partner</td>
<td>466 (6%)</td>
<td>66 (1%)</td>
</tr>
<tr>
<td>Married</td>
<td>3365 (43%)</td>
<td>4088 (53%)</td>
</tr>
<tr>
<td>Separated</td>
<td>99 (1%)</td>
<td>132 (2%)</td>
</tr>
<tr>
<td>Divorced</td>
<td>574 (7%)</td>
<td>450 (6%)</td>
</tr>
<tr>
<td>Widowed</td>
<td>286 (4%)</td>
<td>455 (6%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2204 (28%)</td>
<td>229 (3%)</td>
</tr>
</tbody>
</table>

Discussion

We present an analysis of the determination of patient marital status from both structured and unstructured electronic healthcare data. The present study contributes three novel and significant findings. First, marital status can be obtained from structured data fields; however, in our EHR there is only one field per patient, and the value may not correlate with past encounters due to marital status changes over time. In addition, the structured field is prone to manual entry errors or inconsistent updates when changes occur. Second, we demonstrated a rules-based NLP Engine can extract marital status almost as well as a human reviewer. Third, although there was not a statistically significant difference between NLP and structured field assignment of marital status on a small test set, we saw large discrepancies between the NLP ascertainment of marital status and the structured data fields on a large corpus. This work demonstrated the utility of NLP derived variables in the supplementation of structured electronic healthcare data. For the patients the
NLP Engine was unable to identify a marital status, the distribution of structured labels is shown in Figure 2. The two most common structured labels were Single (39%) and Married (38.5%).

In our cohort, based on the structured data fields, approximately 50% of the patients were married, with single being the next most common status at 25%. The advantage of using structured data is easy for storage and extraction in the electronic healthcare record. However, there remain challenges in the use of structured data fields. First, the marital status definitions can have diverse interpretations, especially culturally sensitive topics in health disparity populations such as racial/ethnic and sexual/gender minorities. Second, the marital status may change over time. This can pose a problem in EHR design, especially if marital status is a patient-level variable. In our institution, marital status is determined at the patient level and changes are only stored in the log databases of the enterprise data warehouse, thus limiting the use of marital status in encounter level analyses. Second, data entry errors may cause multiple entries for marital status. In our analysis, 78% of the changes in the structured marital status fields occurred on the same day. This indicates either data entry errors or conflicting information from different tables in the EHR.

To supplement the limitation of the structured data fields, we developed and validated an NLP engine to ascertain marital status from clinical notes. We elected to limit the notes to clinical Social Work notes as these documents have a high likelihood of containing marital status information. Overall our NLP engine had excellent performance on both a document level (F1 Score 0.97) and patient level (F1 Score 0.95).

Figure 1 – Normalized Comparison of NLP and Structured Data

We compared the performance of our NLP engine with the structured data fields on our full cohort of patients. After analysis of 23,794 notes, the NLP engine was able to make a determination on 72% of notes; however 28% of notes did not contain references to marital status. This large number of unknowns is a limitation of NLP due to the limited information contained in the clinical text. Given we only used Social Work notes, we hypothesize the inclusion of other clinical documentation may decrease the number of unknowns. Eliminating the patients that NLP was unable to make a determination on, we saw good agreement between our NLP engine and the structured data fields (Cohen’s κ 0.6). There was good agreement with married, divorced, widowed, and domestic partner. We saw the most disagreement in patients labeled as Single in the structured data fields. NLP was able to determine those patient are married (10%), domestic partner (14%), divorced (9%). This demonstrated the value of NLP in that sensitive topic such as domestic partners or divorce may be misrepresented in the structured data fields. The largest number of disagreements in each structured data class was related to the NLP engine being unable to ascertain marital status. We analyzed the distribution of structured data labels in those the NLP system was unable to determine marital status.
The distribution of marital status in the unknown patients matched the overall distribution of the structured data fields. This indicates our NLP engine is not biased toward the ascertainment of one particular marital status class.

Figure 2. Structured marital status labels for patient’s without NLP obtained marital status

Overall, based on the results of our study we feel a hybrid approach to the ascertainment of a patient’s marital status should be used. NLP can be used to supplement structured data fields especially in instances where the structured field is Unknown or Single.

Limitations and Future Work

We acknowledge several limitations. First, this study was limited to the traditional definition of marital status as is currently captured in electronic health records. As there is insufficient knowledge of how the spectrum of social/domestic relationships (including same-sex partnerships) are currently documented in the health record, there is an opportunity to extend our work in this direction in the future. The study was limited to one large academic medical center and to one electronic note type (social work notes that were predicted to be of high yield for marital status). These findings merit extension to other note types and other healthcare facilities. Marital status is one of many social risk factors which have become increasingly identified as drivers of health care utilization. Moving forward, we seek to expand out ascertainment of other social risk factors using NLP tools.

Conclusions

In conclusion, we described the determination of marital status from electronic health record data. We developed an NLP engine to inference a patient’s marital status from clinical notes. Overall our NLP engine had excellent performance on both a document level and patient level classification. There was good agreement between the NLP method compared with structured data entries present in the electronic healthcare record. This work demonstrated the value of combining structured and unstructured data for determination of marital status. Our techniques are can be extended to other SDoH and provide researchers valuable insight into the role these factors play in healthcare utilization.

Acknowledgments

This work is funded by a grant from the Agency for Healthcare Research and Quality (K08HS025776) and the VA Salt Lake City Center of Innovation Award (COIN) #I50HX001240. The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The views expressed in this paper are those of the authors and do not necessarily represent the views of the United States Department of Veterans Affairs or the United States government.
References

2018 Salary Survey of AMIA Members: Factors Associated with Higher Salaries

Yan Cheng, PhD\(^1\), April F. Mohanty, MPH, PhD\(^2,3\), Omolola I. Ogunyemi, PhD\(^4,5\), Catherine A. Smith, PhD\(^6\), Gondy Leroy, PhD\(^7\), Qing T. Zeng, PhD\(^1,8\)

\(^1\)Biomedical Informatics Smith, George Washington University, Washington, DC; \(^2\)Informatics Decision-Enhancement Center, Veterans Affairs Medical Center, Salt Lake City, UT; \(^3\)Department of Internal Medicine, School of Medicine, University of Utah, Salt Lake City, UT; \(^4\)Center for Biomedical Informatics, Charles R. Drew University of Medicine and Science, Los Angeles, CA; \(^5\)University of California, Los Angeles, CA; \(^6\)Information School, University of Wisconsin, Madison, WI; \(^7\)Management Information System Department, University of Arizona, Tucson, AZ; \(^8\)Washington DC Veterans Affairs Medical Center, Washington, DC

Abstract

Greater transparency in salaries overall and in factors associated with differing salaries can help students and professionals plan their careers, discover biases and obstacles, and help advance professional disciplines broadly. In March 2018, we conducted the first salary survey of American Medical Informatics Association members. Our goal was to summarize salary information and provide a nuanced view pertaining to the diverse biomedical informatics community. To identify factors associated with higher salaries, we reviewed average salaries for different groups (physician status, academic status, and different leadership positions) by gender. We also fitted multiple linear regression models for all participants (\(N = 201\)) and for gender, physician- and academic-status subgroup. The mean (standard deviation) salary was \$181,774 (\$99,566). Men earned more than women on average, and especially among professionals from academic settings. More years working in informatics and full-time employment were two factors that were consistently associated with higher salary.

INTRODUCTION

Understanding salaries overall and the factors associated with higher salaries in biomedical informatics can aid individuals at all stages of the profession in career planning and negotiation. In addition, greater insight into the characteristics associated with higher compensation in biomedical informatics may help advance the field by identifying potential pay gaps, including gender disparities, which can then be addressed.

In several scientific fields, studies were conducted to explore and discover factors associated with salary.\(^1\)-\(^9\) The majority reported that on average men earned higher salaries than women.\(^2,3,5\)-\(^9\) The American Association of University Women (AAUW) tracks the pay gaps using data from the Bureau of Labor Statistics and the Census Bureau for different characteristics such as race, geographical location and industry. In Fall 2018, they reported a gender pay gap of almost 20% and listed ‘physicians and surgeons’, ‘registered nurses (RNs)’, and ‘medical and health services managers’ among the 10 occupations where the collective loss is the highest for women.\(^10\)

Focused surveys in the medical domain report findings that are in line with the AAUW report. For example, male internists earned 14% more per hour than female internists\(^6\) and male RNs earned \$3,792 to \$17,290 more per year than female RNs, depending on the setting, specialty and position.\(^5\) Among early career physician-researchers, males’ salaries were about \$10,921 higher, after taking other factors into account.\(^3\) Males with a PhD degree earned \$18,000 more annually than females with PhDs.\(^9\) However, a few studies found no gender differences in salary, for example, among radiologists working in public medical schools.\(^1,4\)

The existing work rarely examines factors that may lead to higher salaries for professionals in biomedical informatics and related fields such as health IT. One exception is the 2018 Health Information and Management Systems Society (HIMSS) US Compensation Survey report, which is based on a salary survey of US health IT professionals and conducted annually since 2006. In 2018, responses from 885 participants provided a nuanced view.\(^11\) There were several disparities reported, including lower salaries for women (vs men) and non-white (vs white) professionals in health IT. Compared to males, females earned \$22,797 (accounting for 14%) less on average, and the disparity persisted from 2006 to 2018. Although this study estimated gender and race disparities overall and stratified over several factors separately including age, geographic region, job type, and managerial status; other
factors including educational degree, years of work experience, workplace setting (e.g. industry, academia, government), area(s) of expertise, and professional duties were not examined and may be important to consider.\textsuperscript{3,4,12}

The American Medical Informatics Association (AMIA) is a non-profit professional association dedicated to the development and application of biomedical and health informatics.\textsuperscript{13} Founded in 1989, the association includes over 5,600 informatics professionals.\textsuperscript{14} The members belong to several different groups; they include academic and industry members as well as clinical and non-clinical members. AMIA members also have a wide range of experience levels, ranging from students to Chief Medical Informatics/Information Officers/Chief Health Informatics Officer (CMIO/CHIO) or higher. As such, analysis of this community may provide a diverse and nuanced view on salary and pay differences between different groups. Understanding the range of members' salaries for various informatics roles can inform the AMIA community about existing gaps and may facilitate activities aimed at ensuring fairness and transparency in compensation for all informatics professionals.

The objectives of this paper are to (1) describe the methods of the first AMIA member wide salary survey, (2) describe the distribution of survey participants' salaries overall, (3) describe the factors associated with higher salaries, including a possible gender disparity, and (4) to explore some of the potential barriers to salary and career advancement among AMIA members.

METHODS
Survey Development

In 2017, the Career Advancement Committee ([CAC] formerly named Academic Career Advancement Committee), a subcommittee of the Women in AMIA Steering Committee, created a survey by adapting content from existing surveys to obtain information from AMIA members about their titles, areas of expertise, years in those areas, gender, and salary, in part to identify potential gender-based salary disparities. The CAC collaborated with several members from the Women in AMIA Steering Committee and other AMIA members and staff with expertise in survey development, including reviews of the American Health Information Management Association’s (AHIMA’s) Salary Survey,\textsuperscript{15} to iteratively develop the AMIA Salary Survey. The final AMIA Salary Survey was hosted on SurveyMonkey Inc. and included 34 questions focusing on demographic, education, salary and other forms of compensation, informatics expertise, and current informatics position.

Participant Recruitment and Data Collection

The survey was launched through AMIA eNews on March 22, 2018 as well as two AMIA online communities (the Women in AMIA and Implementation forums) and closed on May 16, 2018. Our goal was to keep the survey open until we reached at least 200 participants. After closing the survey, there were 225 participants. We excluded from the current report those who reported that they did not currently have a salary or were retired (N = 21) or listed a salary of $0 (N = 3).

Statistical Analysis

In descriptive analyses, we present summary statistics overall and statistics stratified by gender, physician (vs non-physician) status, and academic (vs non-academic) status. For analyses stratified by gender we excluded participants who either did not report their gender or reported non-binary/other gender (N=9, 4.5%). For analyses stratified by physician-status we excluded 1 participant who did not report this information. Student’s t-tests and Chi-square tests were conducted to compare mean and proportion difference in areas of informatics expertise and career barriers between two stratified subgroups (i.e., male vs female, etc.).

To estimate the factors influencing salaries of the study cohort, including gender, we fitted a regression model including all participants as well as prediction models for each subgroup of participants, stratified by gender, physician status, and academic status. For each model, we conducted a multiple linear regression with a backward selection procedure, using the zero-skewness logarithm of salaries, as a dependent variable, and all other factors (except for outcome questions such as benefits including health insurance, life insurance, retirement plan, bonus, flex schedule, etc.) collected from the survey as independent variables. To optimize our models, several steps were taken before fitting the multivariate models: 1) if a factor was categorical, we reclassified the original categories in order to get the lowest Akaike Information Criterion (AIC) of that factor in univariate analysis; 2) if factors were highly correlated, we combined them into a single factor (e.g., factors about faculty/faculty rank, similar informatics expertise, etc.); 3) each workplace setting, informatics area, informatics expertise, and career barrier category was modeled as a yes/no indicator variable. The significance level for factor removal from the model was set as 0.1. Multicollinearity was checked using Variance Inflation Factors (VIF) with the rule of thumb that VIF ≥ 5 indicates
If one or more factors had a VIF ≥ 5 in a model, we excluded the factor with the highest VIF and then checked VIF again after exclusion. The process was conducted iteratively, until all factors in the model had a VIF less than 5. The residual normality assumption of each model was checked using plots and tests of skewness and kurtosis. All the analyses were conducted using Stata 13.0.

RESULTS

Overall Participant Characteristics

Among participants who reported a salary greater than $0 (N = 201), more than half were female (62.7%), and physicians and academics were accounted for 38.3% and 59.2%, respectively. Regarding the education of our study sample, the smallest group were participants with a bachelor’s or master’s degree (17.4%), while majority had a Ph.D./DSc (42.8%) or a professional doctorate such as MD, PharmD, and JD (39.3%); and 27.4% reported having two or more doctorate degrees. Among the 201 participants, 91.0% were full-time employees. The number of years of experience as an informatics professional was distributed evenly among the participants, with 17.9% ≤ 4 years, 21.9% 5-9 years, 22.4% 10-14 years, 14.9% 15-19 years, and 22.4% ≥ 20 years. Figure 1 compares the representativeness of the survey participants with AMIA members of 2018. Compared to AMIA members, survey participants had a higher prevalence of females and academics, but a lower prevalence of students. The prevalence of physicians among survey participants was similar to the prevalence in the broader AMIA population.

Figure 1. Demographic characteristics of survey participants and 2018 AMIA members

The lowest salary for someone engaged full-time in informatics activities was $37,000, for a full-time student, while the highest salary reported was $1,900,000. The top two highest salaries ($750,000 and $1,900,000) were identified as outliers (using the empirical rule which identifies outliers beyond 3 standard deviations of the mean of the logarithm of salary). However, we further explored the other self-reported data associated with the top two salaries to determine the plausibility of the high salaries. Based on the other self-reported data (e.g. title, years of experience, and other factors) for the participant who reported the highest salary of $1,900,000, we suspect a data entry error for the salary and used $190,000 instead of $1,900,000 in the data analysis. As for the second outlier, $750,000 could be a true value according to the self-reported data, so we kept it in the data. Tables 1-3 show overviews of the salary data overall and for different groups. The overall mean (standard deviation [SD]) annual salary of the AMIA participants in this study was $181,774 ($99,566) and median (interquartile range) was $165,000 ($111,000-$230,000) (Table 1).

Figure 2 shows that the salary distribution was skewed to the right (p < 0.001), but the zero-skewness logarithm of salary, ln(salary+57,011.99), was normally distributed (p = 1.000).
Table 1. Salaries in US dollars of informatics professionals stratified by gender, physician and academic status

<table>
<thead>
<tr>
<th></th>
<th>Overall (N = 201)</th>
<th>Male (N = 66)</th>
<th>Female (N = 126)</th>
<th>Physician (N = 77)</th>
<th>Non-Physician (N = 123)</th>
<th>Academic (N = 119)</th>
<th>Non-Academic (N = 82)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salary, mean</td>
<td>181,774</td>
<td>206,091</td>
<td>165,369</td>
<td>252,208</td>
<td>138,142</td>
<td>166,618</td>
<td>203,778</td>
</tr>
<tr>
<td>(SD)</td>
<td>(99,566)</td>
<td>(118,865)</td>
<td>(80,807)</td>
<td>(109,448)</td>
<td>(60,930)</td>
<td>(88,579)</td>
<td>(110,545)</td>
</tr>
<tr>
<td>Salary, median</td>
<td>(111,000-)</td>
<td>(125,000-)</td>
<td>(105,000-)</td>
<td>(196,000-)</td>
<td>(98,000-)</td>
<td>(104,000-)</td>
<td>(128,000-)</td>
</tr>
<tr>
<td>(Q1-Q3)</td>
<td>230,000</td>
<td>250,000</td>
<td>215,000</td>
<td>304,000</td>
<td>175,000</td>
<td>220,000</td>
<td>250,000</td>
</tr>
</tbody>
</table>

All unadjusted mean differences (i.e. male vs female, physician vs non-physician, and academic vs non-academic) were statistically significant (p < 0.05). Abbreviations: SD, standard deviation; Q1, the first quartile; Q3, the third quartile.

Table 2. Salaries in US dollars of informatics professionals stratified by physician status and gender

<table>
<thead>
<tr>
<th></th>
<th>Physician (N = 72)</th>
<th>Non-Physician (N = 119)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male (N = 37)</td>
<td>Female (N = 35)</td>
</tr>
<tr>
<td>Salary, mean</td>
<td>259,649 (126,724)</td>
<td>236,514 (80,887)</td>
</tr>
<tr>
<td>(SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salary, difference in means (male vs female)</td>
<td>23,135</td>
<td>-391</td>
</tr>
<tr>
<td>Salary, median</td>
<td>(196,000-310,000)</td>
<td>(200,000-290,000)</td>
</tr>
<tr>
<td>(Q1-Q3)</td>
<td>240,000</td>
<td>235,000</td>
</tr>
</tbody>
</table>

Abbreviations: SD, standard deviation; Q1, the first quartile; Q3, the third quartile.

Table 3. Salaries in US dollars of informatics professionals stratified by academic status and gender

<table>
<thead>
<tr>
<th></th>
<th>Academic (N = 115)</th>
<th>Non-Academic (N = 77)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male (N = 37)</td>
<td>Female (N = 78)</td>
</tr>
<tr>
<td>Salary, mean</td>
<td>199,973 (106,996)</td>
<td>150,070 (73,779)</td>
</tr>
<tr>
<td>(SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salary, difference in means (male vs female)</td>
<td>49,903</td>
<td>23,667</td>
</tr>
<tr>
<td>Salary, median</td>
<td>(120,000-263,000)</td>
<td>(100,000-188,000)</td>
</tr>
<tr>
<td>(Q1-Q3)</td>
<td>205,000</td>
<td>130,000</td>
</tr>
</tbody>
</table>

Abbreviations: SD, standard deviation; Q1, quartile 1; Q3, quartile 3.

Figure 2. Salary and zero-skewness logarithm of salary distribution shown in histogram graphs
Unadjusted Comparisons by Gender

On average, men earned $206,091 (SD: $118,865) per year, compared to $165,369 (SD: $80,807) per year for women, resulting in a gap of $40,722 (Table 1). The largest unadjusted difference in mean salary, where men earned more than women, was for academics ($49,903), followed by non-academics ($23,667) and physicians ($23,135) (Table 2 and 3). However, among non-physicians, men earned slightly less ($391) than women. In general, men and women had a similar number of informatics expertise areas (5.8 vs 5.7). Figure 3 shows the distribution of members who had a leadership position. There were 81 (40.3%) participants who held a leadership role (e.g. CMIO, vice president [VP], director, and department chair). The percentage of leaders was higher among men than women (53.0% vs 34.1%). The top 3 (out of 27) areas of informatics expertise among men were clinical decision support, implementation, and human computer interaction; and among women were clinical decision support, implementation, and clinical research informatics. In comparison, clinical decision support, human computer interaction, and security were more frequently reported as areas of expertise by male participants, while consumer health informatics was more frequently reported by female participants. Figure 4 shows salary comparisons by years of experience. Salaries for both men and women increased with more years in the informatics profession, with the highest mean salary associated with working for 15-19 years among men and working for 10-14 years among women. There were no consistent differences in men versus women across years spent in the profession.

![Figure 3](https://www.amia.org/sites/default/files/AMIA-2018-Salary-Survey-20181031.pdf)

**Figure 3.** Proportion of participants having a leadership position (e.g. CMIO, VP, director, and department chair, N=81)

![Figure 4](https://www.amia.org/sites/default/files/AMIA-2018-Salary-Survey-20181031.pdf)

**Figure 4.** Annual salary by number of years working in the informatics profession stratified by gender (N = 66 males and 126 females)

The survey also requested a list of career advancement barriers. The different rankings of career barriers by gender are shown in Table 4 and rankings by strata of physician and academic status are available online (https://www.amia.org/sites/default/files/AMIA-2018-Salary-Survey-20181031.pdf). Across 5 potential career advancement barriers, on average, men reported fewer than half the number of barriers that women did (0.2 vs 0.6,
p<0.001). The #1 top career barrier for men and women was “other barriers” (i.e., work-life balance, career/life change, lack of support from family) (N=5, 7.6%) and taking time off to raise a family (N=22, 17.5%), respectively.

Qualitative analysis was performed of free-text responses to one question in the salary survey: “Are there significant life events that may have impeded your career progress?” The responses were classified according to the typology proposed by Cachon17 as personal (inherent to the individual); organizational (determined by the organization); and social (determined by sociocultural contexts in which both organization and individual are embedded). Several participants listed multiple factors – “All of the above in one way or another, and at one time or another”.

The free-text responses provide more nuance: Personal: “Two-body problem that comes from being an academic spouse”; “Special Needs Child, had to file FMLA in the past to meet child’s needs”; and “Inability to move to new city due to family commitments.” Organizational: “No family financial support during education; lack of on-the-job mentoring; starting career in areas of the country with lower salaries than where most other positions are”; “Attitudes of tenured faculty”; and “Burnout and compassion fatigue.”

Of the top 5 career barriers (Table 4), three were personal factors, and one simultaneously social and organizational in nature.

Table 4. Top 5 career barriers stratified by gender

<table>
<thead>
<tr>
<th>Barriers number, mean (SD)</th>
<th>Male (N = 66)</th>
<th>Female (N = 126)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top barriers, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.2 (0.5)</td>
<td>0.6 (0.8)</td>
</tr>
<tr>
<td>other barriers</td>
<td>5 (7.6)</td>
<td>taking time off to raise a family</td>
</tr>
<tr>
<td>providing long-term care for loved one</td>
<td>4 (6.1)</td>
<td>other barriers</td>
</tr>
<tr>
<td>taking time off to raise a family</td>
<td>4 (6.1)</td>
<td>providing long-term care for a loved one</td>
</tr>
<tr>
<td>personal illness</td>
<td>2 (3.0)</td>
<td>racial/gender/sexual discrimination in the workplace</td>
</tr>
<tr>
<td>racial/gender/sexual discrimination in the workplace</td>
<td>1 (1.5)</td>
<td>personal illness</td>
</tr>
</tbody>
</table>

Other barriers: i.e., work-life balance, career/life change, lack of supports from family

Unadjusted Comparisons by Physician Status

On average, physicians earned $114,066 per year more than non-physicians (Table 1). Physicians had slightly fewer areas of informatics expertise than non-physicians (5.3 vs 6.0, p=0.152). Table 5 shows the top 3 areas of informatics expertise stratified by physician status. A higher proportion of physicians reported clinician decision support and implementation than non-physicians; while a higher proportion of non-physicians reported data science than physicians. Physicians reported a similar number of career barriers as non-physicians (0.5 vs 0.5, p=0.516).

Table 5. Top 3 areas of informatics expertise stratified by physician status

<table>
<thead>
<tr>
<th>Number of areas of expertise, mean (SD)</th>
<th>Physician (N = 77)</th>
<th>Non-Physician (N = 123)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Areas of Expertise, N (%)</td>
<td>Clinical decision support</td>
<td>Clinical decision support</td>
</tr>
<tr>
<td>1</td>
<td>63 (81.8)</td>
<td>62 (50.4)</td>
</tr>
<tr>
<td>Clinical decision support</td>
<td>Implementation</td>
<td>Implementation</td>
</tr>
<tr>
<td>2</td>
<td>43 (55.8)</td>
<td>49 (39.8)</td>
</tr>
<tr>
<td>Implementation</td>
<td>26 (33.8)</td>
<td>Data science</td>
</tr>
<tr>
<td>3</td>
<td>Human computer interaction</td>
<td>45 (36.6)</td>
</tr>
</tbody>
</table>
Unadjusted Comparisons by Academic Status

Academics earned $37,160 per year less than non-academics (Table 1). Academics had more areas of informatics expertise than non-academics (6.1 vs 5.1, p=0.022). Table 6 shows the top 3 areas of informatics expertise stratified by academic status. A greater proportion of academics than non-academic counterparts reported clinical research informatics; while more non-academics reported standards. Academics identified a similar number of career barriers as non-academics (0.5 vs 0.5, p=0.899).

Table 6. Top 3 areas of informatics expertise stratified by academic status

<table>
<thead>
<tr>
<th></th>
<th>Academic (N = 119)</th>
<th>Non-Academic (N = 82)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of areas of expertise, mean (SD)</td>
<td>6.1 (3.5)</td>
<td>5.1 (2.8)</td>
</tr>
<tr>
<td>Top Areas of Expertise, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1  Clinical decision support</td>
<td>71 (59.7)</td>
<td>55 (67.1)</td>
</tr>
<tr>
<td>2  Implementation</td>
<td>50 (42.0)</td>
<td>42 (51.2)</td>
</tr>
<tr>
<td>3  Clinical research informatics</td>
<td>49 (41.2)</td>
<td>29 (35.4)</td>
</tr>
</tbody>
</table>

Fully Adjusted Analysis

We fitted a model to identify the factors independently associated with the zero-skewness logarithm of salary. As shown in Table 7, gender was not a statistically significant factor associated with salary and was removed from the model. The factors that were associated with higher salaries, and were statistically significant, included physician (vs non-physician), more years working as an informatics profession, full- (vs part-) time employment, a higher number of people directly reporting to participant, having a leadership position (vs non-leadership), hospital/health system (vs non-hospital/health system), expertise of consumer health informatics, and natural language processing. The factors that were associated with lower salaries, and were statistically significant, included having a combination of experience and education that was not part of informatics training, public health informatics area, and mobile health expertise. Also, “other career barriers” (i.e., work-life balance, career/life change, lack of support from family) were associated with lower salaries in the fully adjusted model. For example, the coefficient of physician was 0.312, which means if the annual salary of both physicians’ and non-physicians’ was increased by $57,012, physicians’ salary would be 1.366 times as non-physicians’ on average, after controlling for other factors. The R-squared of the model was 0.674, which means 67.4% of variance in the dependent variable (zero-skewness logarithm of salary) was explained by the factors included in the model.

In Figure 5, all plots showed that the residuals of the predictive model were normally distributed. The p-value of the skewness and kurtosis test for normality of the residuals was 0.0861, so we could not reject the hypothesis that the residuals were normally distributed.

Figure 5. Normality of residual diagnosis by residual vs. fitted plot, Q-Q plot, and histogram of the residuals
We also conducted a sensitivity analysis after excluding the outlier (extremely high salary of $750,000). All significant factors in the model remained statistically significant and estimates of predictors were in same direction, with the predictor magnitudes just slightly changed. Two more statistically significant factors were identified after excluding the outlier: education degree (higher degree associated with higher salary) and the career barrier of “Taking time off to raise a family” (barrier associated with lower salary).

Table 7. Predictive model of zero-skewness logarithm of salary among overall participants

<table>
<thead>
<tr>
<th>Factors predicting ln(salary + 57,011.99)</th>
<th>β</th>
<th>p-value</th>
<th>95% CI of β</th>
<th>Exponential of β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician vs non-physician</td>
<td>0.312</td>
<td>&lt;0.001</td>
<td>0.226 - 0.380</td>
<td>1.366</td>
</tr>
<tr>
<td>Prior career-field or an educational background that was not part of informatics training</td>
<td>0.000</td>
<td>-</td>
<td>-</td>
<td>1.000</td>
</tr>
<tr>
<td>“No” / “Yes, education only” / “Yes, experience only”</td>
<td>0.014</td>
<td>&lt;0.001</td>
<td>-0.215 - 0.066</td>
<td>0.868</td>
</tr>
<tr>
<td>“Yes, a combination of experience and education”</td>
<td>-0.141</td>
<td>&lt;0.001</td>
<td>-1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Years working as an informatics professional</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤4 Years</td>
<td>0.000</td>
<td>-</td>
<td>-</td>
<td>1.000</td>
</tr>
<tr>
<td>5-9 Years</td>
<td>0.286</td>
<td>&lt;0.001</td>
<td>0.176 - 0.395</td>
<td>1.331</td>
</tr>
<tr>
<td>≥10 Years</td>
<td>0.333</td>
<td>&lt;0.001</td>
<td>0.237 - 0.428</td>
<td>1.395</td>
</tr>
<tr>
<td>Full time vs part time</td>
<td>0.428</td>
<td>&lt;0.001</td>
<td>0.297 - 0.559</td>
<td>1.534</td>
</tr>
<tr>
<td>Number of people directly report to participant</td>
<td>0.001</td>
<td>0.040</td>
<td>0.001 - 0.002</td>
<td>1.001</td>
</tr>
<tr>
<td>Leadership position vs no leadership position</td>
<td>0.176</td>
<td>&lt;0.001</td>
<td>0.095 - 0.257</td>
<td>1.192</td>
</tr>
<tr>
<td>Place of employment gives a cost of living increase</td>
<td>0.069</td>
<td>0.075</td>
<td>-0.007 - 0.145</td>
<td>1.071</td>
</tr>
<tr>
<td>Workplace setting of Hospital/Health system</td>
<td>0.076</td>
<td>0.043</td>
<td>0.002 - 0.149</td>
<td>1.079</td>
</tr>
<tr>
<td>Area of informatics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public health informatics</td>
<td>-0.099</td>
<td>0.018</td>
<td>-0.181 - -0.017</td>
<td>0.906</td>
</tr>
<tr>
<td>Expertise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer health informatics</td>
<td>0.147</td>
<td>0.002</td>
<td>0.057 - 0.237</td>
<td>1.158</td>
</tr>
<tr>
<td>Mobile health</td>
<td>-0.099</td>
<td>0.031</td>
<td>-0.191 - -0.008</td>
<td>0.906</td>
</tr>
<tr>
<td>Natural language processing</td>
<td>0.156</td>
<td>0.001</td>
<td>0.063 - 0.248</td>
<td>1.169</td>
</tr>
<tr>
<td>Barrier</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others (i.e., work-life balance, career/life change, lack of support from family)</td>
<td>-0.132</td>
<td>0.015</td>
<td>-0.237 - -0.026</td>
<td>0.876</td>
</tr>
</tbody>
</table>

Fully Adjusted Analysis by Strata of Physician, Non-Physician, Academic, and Non-Academic subgroup

We also fitted models for each subgroup and found that the statistically significant factors varied in the different models. Gender was only an independent factor associated with salary for the model among academics (N = 115). In academics, the mean salary for men was higher than the mean salary for women (β=0.302, p=0.008) after controlling for other factors. Across all models of subgroups, there were only two factors that were consistently associated with higher salary and were statistically significant: 1) more years working as an informatics professional and 2) full-time (vs part-time) employment.

DISCUSSION

To the best of our knowledge, this is the first study to report on the salaries of AMIA members. Overall the mean salary of the 201 salary survey participants was $181,774. Further, our results revealed that overall, men earned a higher salary than women on average. However, after adjusting for other factors, gender was not associated with salary in participants overall or by subgroups of academic and physician status strata, except among participants who reported an academic position, where men earned higher salaries on average. Other characteristics associated with salary may explain the differences across gender. Men (vs women) had a higher education level (doctorate level) (87.9% vs 78.6%), a higher proportion of physicians (56.1% vs 27.8%) and leaders (53.0% vs 34.1%). A lower proportion of men reported career barriers than women (22.7% vs 46.8%). These initial findings should be interpreted with caution, noting that 40.3% of participants reported that they held a higher level leadership position and may over-represent professionals who earn much more than those in non-leadership positions.

Overall, the participants in the AMIA salary survey reported a mean salary between the mean salaries reported by the 2018 HIMSS US Compensation Survey ($109,610) and that reported by the 2018 Association of Medical
Directors of Information Systems-Gartner survey for CMIO/CHIOs ($353,600).\textsuperscript{18,19} The mean salary of AMIA members is also in line with those reported by the Association of American Medical College’s Faculty Salary Report, Fiscal Year 2018 of $184,300 among MDs in the basic sciences or $145,900 among PhDs in the clinical sciences and point to the potentially lucrative nature of the biomedical informatics profession.\textsuperscript{20} On the other hand, the mean salary among academic AMIA salary survey participants ($166,618) was lower vs biomedical informatics faculty in the 2018 Association of American Medical Colleges Faculty Salary Report ($246,500).\textsuperscript{21}

Several of the differences observed in salaries among participants in our study were consistent with the HIMSS Compensation Survey.\textsuperscript{18} As in our study, in HIMSS\textsuperscript{19}, men on average earned more than women, $123,244 versus $100,477 per year, respectively. Findings from the HIMSS survey\textsuperscript{18} also suggested that salaries increased with age. In our study, we used years working in the informatics profession as a surrogate of age. We found that working for 5 years or more (versus fewer than 4 years) was associated with higher salaries and was statistically significant. However, salary increased continuously with more years worked but only among non-physicians. Unlike the HIMSS survey, where the gender salary gap widened with older age, the gender gap in our study remained relatively constant with increased years working in the biomedical informatics profession (results not shown).

Consistent with previous research, we observed that there were more barriers and different ranking of top career barriers across men and women, which may partly account for the higher salaries observed among men and higher percentage of men in leadership positions compared to women among participants overall.\textsuperscript{22,23} Whether differences in career burden due to taking time off to raise a family (#1 barrier for women) versus “other barriers” (i.e. work-life balance, career/life change, lack of support from family, #1 barrier for men) help explain the gender pay gap, as well as differences in resources available to address these barriers for men versus women, requires further research.

The findings of this study should be interpreted with caution and in the context of several limitations. First, the sample size is limited, especially for males. By comparison, the HIMSS study had 885 participants, almost 4 times the number of participants for ours, at 225. The gender ratio of women to men in this study was 2:1. Another limitation is that to protect participant identity, we did not collect sociodemographic factors such as age, race, and ethnicity. A third limitation is that in this study, we had more responses from individuals with senior or leadership-level positions than early-career individuals. For example, there were 20 full professors vs 17 assistant professors, and 81 individuals holding leadership roles vs 2 post-doctoral fellows. Very few students (N=2, 1.0%) and postdocs/fellows (N=8, 4.0%) participated in the survey. In our sample, 34% of women held leadership positions. On average, the number of women holding the most senior-level positions (such as chief-level positions) has not increased since 2015 and remains at 22%,\textsuperscript{24} even lower representation is reported for specific roles such as hospital CEOs (18% female) and deans and department chairs (16%).\textsuperscript{25} Due to the limited sample size, we did not evaluate the interaction effect between these factors and gender on salary levels.

That said, this study represents a promising start in identifying factors associated with higher salaries among informatics professionals. We intend to repeat this salary survey annually or biannually, with AMIA’s permission and assistance. We aim to increase and improve representation by reaching out to the full spectrum of informatics professionals for different genders, ages, races/ethnicities, and other characteristics where there may be gaps in compensation. Also, with a larger sample size, we will be able to take into account additional variables such as race and ethnicity and perform time-series analysis and study salary trends.

CONCLUSION

Overall, our study builds on the prior work that suggests that field of biomedical informatics may be a lucrative career choice for many in terms of salary. Gender disparity was only observed for professionals who reported an academic work setting but not in other groups (i.e. physician, non-physician, and non-academic subgroups), when specific factors, not previously considered by prior related studies, were included in the models. More years working as an informatics professional and full-time (vs part-time) employment were the only two factors that were consistently associated with higher salary across all groups and remained statistically significant. Future salary surveys of AMIA members will seek to address the limitations of the current study including small sample size and high proportion of the participants holding leadership roles, to improve representation and the generalizability of the findings and to continue to advance the biomedical informatics profession.

Acknowledgments: The authors would like to thank Megan Baldwin and Theresa Jones for their assistance with survey development and Wendy Chapman for her assistance in recruiting them to participate in survey development activities.
References

Serial Heart Rate Variability Measures for Risk Prediction of Septic Patients in the Emergency Department

Calvin J. Chiew, MBBS, MPH1*, Han Wang, BEng2*,
Marcus E. H. Ong, MBBS, MPH3,4,5, Ting Hway Wong, MB BChir, MPH6, Zhi Xiong Koh, BEng5, Nan Liu, PhD3,4^, Mengling Feng, PhD2^

1Preventive Medicine Residency Program, National University Health System, Singapore;
2Saw Swee Hock School of Public Health, National University Health System, National University of Singapore, Singapore;
3Health Services Research Centre, Singapore Health Services, Singapore;
4Health Services and Systems Research, Duke-NUS Medical School, National University of Singapore, Singapore;
5Department of Emergency Medicine, Singapore General Hospital, Singapore;
6Department of General Surgery, Singapore General Hospital, Singapore;

Abstract

In this study, we used serial heart rate variability (HRV) measures over 2 hours to improve the prediction of 30-day in-hospital mortality among septic patients in the emergency department (ED). We presented a generalizable methodology for processing and analysing HRV time series (HRVTS) data which may be noisy and incomplete. Feature sets were created from the HRVTS data of 162 patients with suspected sepsis using aggregation-based, delta-based and regression-based series-to-point transformations, and modelled over 100 random stratified splits. An optimized feature set comprising 12 selected HRVTS features performed better than baseline feature sets which only included patient demographics, vital signs and single time-point HRV measures taken at triage. This improved risk stratification approach could be used in the ED to identify high-risk septic patients for appropriate management and disposition.

Introduction

Sepsis is an increasingly recognised global problem that has a 10-20% in-hospital mortality (IHM) rate1-3. Risk stratification of septic patients in the Emergency Department (ED) may help to guide appropriate management and disposition, thereby reducing morbidity and mortality4-6. A number of clinical tools, such as the Quick Sequential Organ Failure Assessment (qSOFA), National Early Warning Score (NEWS) and Modified Early Warning Score (MEWS), have been used to risk stratify septic patients in the ED, where certain clinical information, such as laboratory investigations, are initially not available. These risk calculators rely solely on patient demographics, vital signs and clinical observations in their scoring criteria.

Several studies have reported the prognostic value of heart rate variability (HRV) parameters in septic patients presenting to the ED7,9. Septic patients have reduced sympatho-vagal balance and impaired sympathetic activity, which lead to varying degrees of cardiac autonomic dysfunction10. This can be detected by HRV analysis, a quick, non-invasive technique of evaluating the beat-to-beat variation in heart rate. HRV analyses are divided into linear and non-linear methods11. Linear methods include HRV parameters measured in time or frequency domains. Time domain HRV parameters are statistical calculations of consecutive R-R time intervals and how they correlate with each other. Frequency domain HRV parameters are based on spectral analysis. Studies have suggested that regulators of the cardiovascular system interact in a non-linear way12, 13 and HRV analysis using non-linear methods reflect these mechanisms14.

We previously demonstrated that HRV parameters measured at triage, combined with patient demographics and vital signs, can improve the prediction of 30-day IHM among septic patients in the ED compared to traditional risk stratification tools15, 16. Since sicker patients are often put on continuous electrocardiogram (ECG) monitoring in the
ED, their serial HRV measurements over time can also be derived. We postulated that HRV time series (HRVTS) may embed more prognostic information than HRV parameters derived at a single time-point, such as at triage.

To date, there has been no published study exploring the use of HRVTS for risk prediction in sepsis. Thus, in this study, we aimed to compare the performance of HRVTS over 2 hours against single time-point HRV measures at triage, in the prediction of 30-day IHM among adult septic patients in the ED setting. In addition, since serial HRV measures are seldom encountered in literature, we aimed to develop and present a generalizable methodology for processing and analysing such data which may be noisy and incomplete.

**Methods**

**Data preparation**

Ethics approval for the study was obtained from SingHealth’s Centralised Institutional Review Board (CIRB, Reference Number 2016/2858), with waiver of patient consent. Patients above 21 years old who presented to the Singapore General Hospital (SGH) ED between September 2014 and April 2016 with clinically suspected sepsis, and who met at least 2 of the 4 Systemic Inflammatory Response Syndrome (SIRS) criteria\(^1\) were included. The SIRS criteria are temperature (<36°C or >38°C), heart rate (>90 beats/min), respiratory rate (>20 breaths/min) and total white count (<4000/mm\(^3\) or >12000/mm\(^3\)).

Patient demographics (age, gender, ethnicity), first set of vital signs recorded at triage (temperature, respiratory rate, heart rate, systolic and diastolic blood pressures, and Glasgow Coma Scale (GCS) score), and outcome of interest (IHM within 30 days of ED admission) were retrieved from the hospital’s electronic medical records. Continuous one-lead ECG tracings over 2 hours in the ED were also obtained from X-Series Monitor (ZOLL Medical Corporation, Chelmsford, MA). The ECGs were split into 8 intervals of 15 minutes each and then analysed for HRV parameters using the most stable continuous 5-minute segment within each interval. 22 HRV measures in time, frequency and non-linear domains were computed for each interval. For patients whose ECGs were shorter than 2 hours or unanalysable due to non-sinus rhythm, excessive artifacts and/or ectopic beats, their data was considered missing for the relevant intervals.

To ensure data quality and to limit the amount of missing data in subsequent analysis, we only included patients with at least 7 of the 8 intervals in our final analysis (i.e. at most 1 missing interval allowed). Univariate statistical analysis was conducted to compare patients with and without outcome, as well as to compare the original and final cohorts (Kruskal-Wallis H test for continuous variables and Chi-squared test for categorical variables).

**Feature Extraction**

We created 2 baseline feature sets that did not utilize any information from the HRVTS as baselines for comparison. We then designed 3 groups of candidate feature sets (aggregation-based, delta-based and regression-based) using different methods of series-to-point transformations to encapsulate the HRVTS information. Detailed definitions of the feature sets, as well as an example of how the feature sets were calculated, are presented in Tables 1 and 2 respectively.

**Baseline feature sets:** The first baseline feature set (fs_no_HRV) contained only patient demographics and vital signs. These features are included in all subsequent feature sets. In the second baseline feature set (fs_baseline), single time-point HRV parameters generated from the first 5-minutes of ECG obtained at triage were added.

**Aggregation-based features sets:** 3 aggregation-based feature sets were constructed by taking (1) the first and last values of the HRVTS (fs_first_last), (2) the median and interquartile range (IQR) of the HRVTS (fs_med_IQR), and (3) the minimum and maximum values of the HRVTS (fs_min_max).

**Delta-based feature sets:** With the assumption that the magnitude of relative changes between neighbouring intervals were informative, we calculated (1) the first and last (fs_delta_first_last), (2) the median and IQR (fs_delta_med_IQR), and (3) the minimum and maximum (fs_delta_min_max) of the absolute differences between adjacent HRVTS measurements. Linear interpolation was used to impute missing measurements.

**Regression-based feature sets:** With the assumption that trends in the HRVTS were informative, we fitted the (1) original values (fs_reg_default), (2) absolute differences between measurements (fs_reg_delta), and (3) cumulative differences between measurements (fs_reg_cum_delta) of the HRVTS to a linear regression model.
using the ordinary least squares method. The linear coefficients and intercepts obtained were used as features to characterize the trends.

Table 1. Definitions of the feature sets with feature values calculated based on the example mean \textit{rr} HRVTS.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature set</th>
<th>Definition of features (additional to \texttt{fs_no_HRV})</th>
<th>Name of HRV features</th>
<th>Feature value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>\texttt{fs_no_HRV}</td>
<td>Only gender, race, age and the vital signs</td>
<td>mean\textit{rr}</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>\texttt{fs_baseline}</td>
<td>First 5-min HRV parameter at triage</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Aggregation</td>
<td>\texttt{fs_first_last}</td>
<td>First and last value of HRVTS</td>
<td>first\textit{mean_rr}</td>
<td>514.00</td>
</tr>
<tr>
<td></td>
<td>\texttt{fs_med_IQR}</td>
<td>Median and inter-quantile range of HRVTS</td>
<td>median\textit{mean_rr}</td>
<td>501.00</td>
</tr>
<tr>
<td></td>
<td>\texttt{fs_min_max}</td>
<td>Minimum and maximum of HRVTS</td>
<td>min\textit{mean_rr}</td>
<td>406.00</td>
</tr>
<tr>
<td>Delta</td>
<td>\texttt{fs_delta_first_last}</td>
<td>First and last of all the differences in absolute value between neighbouring intervals</td>
<td>delta\textit{first_mean_rr}</td>
<td>58.00</td>
</tr>
<tr>
<td></td>
<td>\texttt{fs_delta_med_IQR}</td>
<td>Median and inter-quantile range of all the differences in absolute value between neighbouring intervals</td>
<td>delta\textit{median_mean_rr}</td>
<td>78.00</td>
</tr>
<tr>
<td></td>
<td>\texttt{fs_delta_min_max}</td>
<td>Minimum and maximum of all the differences in absolute value between neighbouring intervals</td>
<td>delta\textit{min_mean_rr}</td>
<td>50.00</td>
</tr>
<tr>
<td>Regression</td>
<td>\texttt{fs_reg_default}</td>
<td>Coefficient and intercept of the linear regression model fit on the HRVTS</td>
<td>coef\textit{mean_rr}</td>
<td>10.86</td>
</tr>
<tr>
<td></td>
<td>\texttt{fs_reg_delta}</td>
<td>Coefficient and intercept of the linear regression model fit on the series of the differences in absolute value between neighbouring intervals</td>
<td>coef\textit{delta_mean_rr}</td>
<td>2.40</td>
</tr>
<tr>
<td></td>
<td>\texttt{fs_reg_cum_delta}</td>
<td>Coefficient and intercept of the linear regression model fit on the series of cumulative sum of all the differences in absolute value between neighbouring intervals</td>
<td>coef\textit{cum_delta_mean_rr}</td>
<td>91.00</td>
</tr>
</tbody>
</table>

Table 2. Example of a patient’s mean \textit{rr} HRVTS and derived values for the calculation of features. The 4th interval was missing, and the interpolated value is indicated by an underline.

<table>
<thead>
<tr>
<th>Time</th>
<th>mean\textit{rr}</th>
<th>Derived from raw value</th>
<th>Absolute \textit{delta}</th>
<th>Derived from absolute \textit{delta}</th>
<th>Cumulative absolute \textit{delta}</th>
<th>Derived from cumulative absolute \textit{delta}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>514.00</td>
<td>median</td>
<td>58.00</td>
<td>median\textit{delta}</td>
<td>58.00</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>456.00</td>
<td>median</td>
<td>50.00</td>
<td>median 78.00</td>
<td>108.00</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>406.00</td>
<td>IQR</td>
<td>113.00</td>
<td>IQR 57.25</td>
<td>221.00</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>missing</td>
<td>IQR</td>
<td>113.00</td>
<td>IQR</td>
<td>334.00</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>632.00</td>
<td>Linear \textit{y} = -10.86x</td>
<td>98.00</td>
<td>Linear \textit{y} = -2.40x</td>
<td>432.00</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>534.00</td>
<td>Regression + 463.57</td>
<td>46.00</td>
<td>Regression + 71.27</td>
<td>478.00</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>488.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

\textit{(HRVTS, heart rate variability time series; mean\textit{rr}, average width of the RR interval)}

Model Setup

Prior to modelling, all features were standardized to be between 0 to 1 by MinMaxScaler (subtracting the minimum value and then dividing by the difference between the maximum and minimum values).

We set aside 20% of the patients as test data, stratified by outcome. Within the remaining 80% of the patients, we created 100 random stratified splits, with 80% as training data and 20% as validation data in each split. We chose logistic regression with L1 regularization (LASSO) with C=1.0 as our model due to its simplicity and the interpretability of odds ratios obtained. Class imbalance was addressed by applying sample weights inversely proportional to the number of samples. We used Area Under the Receiver Operating Characteristic curve (AUC) as our performance evaluation metric.
**Feature Selection**

Each feature set was inputted separately into the logistic regression model and the median odds ratios of all non-zero coefficients obtained across 100 splits were calculated. We tested the proportion of each feature having non-zero coefficient larger than 0.5 using one-sided binomial test under Bonferroni correction.

To create an optimized feature set, the feature with the most extreme odds ratio among statistically significant ones (if any) was selected for each HRV parameter. These selected HRV features were then combined with demographic and vital sign variables significant across majority of feature sets to form an optimized feature set `fs_final_selection`.

HRV analysis was performed using Kubios HRV software version 2.2 (Kuopio, Finland)\(^\text{18}\). Statistical analysis and modelling were carried out in Python 3.6 (Python Software Foundation, Wilmington, Delaware, USA) using the statsmodel\(^\text{19}\) and scikit-learn libraries\(^\text{20}\).

**Results**

Figure 1 shows the cohort selection process. 214 patients were in the original cohort whose ECGs were split into intervals and analysed for HRV parameters. After excluding patients with more than 1 missing interval, 162 patients were included in the final cohort, of whom 31 (19.1%) met the outcome. There were no significant differences in the patient demographics, vital signs and outcome between the original and final cohorts (Table 3).

![Cohort selection flowchart](image)

**Figure 1.** Cohort selection flowchart.

Among the final cohort, 77 (47.5%) of the patients were male, with median age of 67 years (inter-quartile range, IQR 57–79). Patients who met the outcome were older (median age 76 years; IQR 68–83 years) than those who did not (median age 66 years; IQR 57–77 years). There were no significant differences in gender and ethnicity distributions between the two groups. In terms of vital signs, patients who met the outcome had higher respiratory rates, as well as lower temperatures, heart rates, systolic blood pressures and GCS scores, compared to patients who did not meet the outcome. The HRVTS of the two groups are visualised in Figure 2.

Table 4 shows the median odds ratios across all feature sets and predictor variables. Most of the features had coefficients of zero (hence odds ratio of zero) due to the L1 regularization. 12 HRV parameters (5 in time domain, 4 in frequency domain and 3 in non-linear domain) had statistically significant odds ratios and were selected into the optimized feature set `fs_final_selection` (Table 5). 3 other non-HRV variables (age, systolic blood pressure and GCS) were also statistically significant across most feature sets and added to the optimized feature set `fs_final_selection` to make a total of 15 features.
Table 3. Patient demographics and vital signs of the original and final cohorts used in the analysis.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Original Cohort</th>
<th>Final Cohort</th>
<th>Original vs. Final</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No 30-day IHM</td>
<td>30-day IHM</td>
<td>P-Value</td>
</tr>
<tr>
<td>n (%)</td>
<td>174 (81.3)</td>
<td>40 (18.7)</td>
<td>-</td>
</tr>
<tr>
<td>Age (years)</td>
<td>66.0 [56.2, 76.8]</td>
<td>76.0 [67.8, 83.0]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Male gender</td>
<td>88 (50.6)</td>
<td>20 (50.0)</td>
<td>0.913</td>
</tr>
<tr>
<td>Ethnicity (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>125 (71.8)</td>
<td>30 (75.0)</td>
<td>Ref.</td>
</tr>
<tr>
<td>Malay</td>
<td>25 (14.4)</td>
<td>4 (10.0)</td>
<td>0.637</td>
</tr>
<tr>
<td>Indian</td>
<td>15 (8.6)</td>
<td>5 (12.5)</td>
<td>0.545</td>
</tr>
<tr>
<td>Others</td>
<td>9 (5.2)</td>
<td>1 (2.5)</td>
<td>0.692</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>38.3 [37.4, 38.8]</td>
<td>37.4 [36.5, 38.5]</td>
<td>0.005</td>
</tr>
<tr>
<td>Respiratory rate (breaths/ min)</td>
<td>19.0 [18.0, 22.0]</td>
<td>22.0 [18.8, 26.0]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Heart rate (beats/min)</td>
<td>118.0 [105.0, 130.0]</td>
<td>108.5 [97.8, 125.5]</td>
<td>0.078</td>
</tr>
<tr>
<td>Systolic BP (mmHg)</td>
<td>114.0 [91.0, 142.8]</td>
<td>105.0 [91.0, 126.2]</td>
<td>0.097</td>
</tr>
<tr>
<td>Diastolic BP (mmHg)</td>
<td>63.0 [55.0, 74.0]</td>
<td>61.5 [52.0, 70.2]</td>
<td>0.442</td>
</tr>
<tr>
<td>GCS score</td>
<td>15.0 [13.2, 15.0]</td>
<td>14.0 [9.0, 15.0]</td>
<td>0.002</td>
</tr>
</tbody>
</table>

(BP, blood pressure; GCS, Glasgow Coma Scale)

For continuous variables, data is presented in medians and interquartile ranges. Kruskal-Wallis H test was used to test for differences. For categorical variables, data is presented in frequencies and percentages. Chi-square test was used to test for association.

Figure 2. Visualization of HRVTS for patients with outcome (orange) and without outcome (blue). (a) Time domain parameters (b) Non-linear domain parameters (c) Frequency domain parameters.
Table 4. Median odds ratio of all the features. Green cells indicate odds ratio > 1, and red cells indicate odds ratio < 1. The darker the color, the stronger the odds ratio deviates from 1. All significant features are in black text. Values with a blue box are selected for fs_final_selection.

<table>
<thead>
<tr>
<th>Item</th>
<th>no HRV</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean RR (s)</td>
<td>1.48</td>
<td>1.52</td>
</tr>
<tr>
<td>SD RR (s)</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>Mean HR (bpm)</td>
<td>5.33</td>
<td>6.18</td>
</tr>
<tr>
<td>SD HR (bpm)</td>
<td>0.33</td>
<td>0.50</td>
</tr>
<tr>
<td>RMSSD (s)</td>
<td>3.40</td>
<td>3.00</td>
</tr>
<tr>
<td>NNS50 count</td>
<td>3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>pNN50 (%)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>RR triangular index</td>
<td>6.03</td>
<td>2.21</td>
</tr>
<tr>
<td>Frequency Domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total power (m2)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>VLF power (m2)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>LF power (m2)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HF power (m2)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>LF power norm (m.u)</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>HF power norm (m.u)</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>Trend</td>
<td>2.75</td>
<td>4.26</td>
</tr>
<tr>
<td>Non-linear Domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poincare plot SD1 (ms)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Poincare plot SD2 (ms)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Approximate entropy</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>Sample entropy</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>DFA a-1</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>DFA a-2</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>2.18</td>
<td>1.91</td>
</tr>
<tr>
<td>Gender (male)</td>
<td>1.55</td>
<td>1.55</td>
</tr>
<tr>
<td>Race (Haley)</td>
<td>1.48</td>
<td>1.48</td>
</tr>
<tr>
<td>Race (Indian)</td>
<td>3.05</td>
<td>3.05</td>
</tr>
<tr>
<td>Race (Others)</td>
<td>1.27</td>
<td>1.48</td>
</tr>
<tr>
<td>Vital Signs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>0.27</td>
<td>0.24</td>
</tr>
<tr>
<td>Heart rate (beats/min)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Respiratory rate (breaths/min)</td>
<td>19.83</td>
<td>23.40</td>
</tr>
<tr>
<td>Systolic BP (mmHg)</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td>Diastolic BP (mmHg)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>GCS score</td>
<td>0.44</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Mean RR, average width of the RR interval; SD RR, standard deviation of all RR intervals; HR, heart rate; RMSSD, root mean square of differences between adjacent RR intervals; NNS50, number of consecutive RR intervals differing by more than 50 ms; pNN50, percentage of consecutive RR intervals differing by more than 50 ms; TINN, baseline width of a triangle fit into the RR interval histogram using a least squares; VLF, very low frequency; LF, low frequency; HF, high frequency; norm, normalized; LF/HF, ratio of LF power to HF power; DFA, detrended fluctuation analysis; BP, blood pressure; GCS, Glasgow Coma Scale.
Table 5. Features in the optimized feature set $fs_{\text{final selection}}$.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Featurization</th>
<th>From Feature Set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time Domain</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean RR (s)</td>
<td>Minimum</td>
<td>$fs_{\text{min max}}$</td>
</tr>
<tr>
<td>Mean HR (bpm)</td>
<td>Intercept</td>
<td>$fs_{\text{reg default}}$</td>
</tr>
<tr>
<td>NN50 (count)</td>
<td>Last</td>
<td>$fs_{\text{delta first last}}$</td>
</tr>
<tr>
<td>RR triangular index</td>
<td>Median</td>
<td>$fs_{\text{delta med IQR}}$</td>
</tr>
<tr>
<td>TINN</td>
<td>Last</td>
<td>$fs_{\text{first last}}$</td>
</tr>
<tr>
<td><strong>Frequency Domain</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VLF power (ms2)</td>
<td>Minimum</td>
<td>$fs_{\text{delta min max}}$</td>
</tr>
<tr>
<td>LF power norm (n.u.)</td>
<td>Intercept</td>
<td>$fs_{\text{reg delta}}$</td>
</tr>
<tr>
<td>HF power norm (n.u.)</td>
<td>Minimum</td>
<td>$fs_{\text{min max}}$</td>
</tr>
<tr>
<td>LF/HF</td>
<td>Intercept</td>
<td>$fs_{\text{reg delta}}$</td>
</tr>
<tr>
<td><strong>Non-linear Domain</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximate entropy</td>
<td>Coef</td>
<td>$fs_{\text{reg cum delta}}$</td>
</tr>
<tr>
<td>Sample entropy</td>
<td>IQR</td>
<td>$fs_{\text{delta med IQR}}$</td>
</tr>
<tr>
<td>DFA, $\alpha$-1</td>
<td>Intercept</td>
<td>$fs_{\text{reg delta}}$</td>
</tr>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>Original</td>
<td>all</td>
</tr>
<tr>
<td>Systolic BP (mmHg)</td>
<td>Original</td>
<td>all</td>
</tr>
<tr>
<td>GCS score</td>
<td>Original</td>
<td>all</td>
</tr>
</tbody>
</table>

Figure 3 summarizes the median AUC of all feature sets over 100 training splits. The optimized HRVTS feature set $fs_{\text{final selection}}$ achieved the highest median AUC of 0.82, followed by the candidate feature set $fs_{\text{min max}}$ which achieved median AUC of 0.77. Both performed better than baseline feature sets which do not contain HRVTS information, namely $fs_{\text{no HRV}}$ (median AUC 0.68) and $fs_{\text{baseline}}$ (median AUC 0.70). Figure 4 compares the median Receiver Operating Characteristic (ROC) curves of these four feature sets over 100 training splits. Table 6 reports the AUC achieved on the test set by these four feature sets. $fs_{\text{final selection}}$ outperformed the rest by a significant margin, while $fs_{\text{min max}}$ did not perform well, suggesting potential overfitting with single featurization method.

Figure 3. Median AUCs of all feature sets over 100 training splits.
**Figure 4.** Receiver Operating Characteristics (ROC) curves of fs_final_selection, fs_min_max and two baseline feature sets. AUCs reported in median ± IQR over 100 training splits.

**Table 6.** Summary of median AUC achieved on validation sets and AUC achieved on test set.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Median AUC (Validation)</th>
<th>AUC (Test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fs_no_HRV</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>fs_baseline</td>
<td>0.70</td>
<td>0.67</td>
</tr>
<tr>
<td>fs_min_max</td>
<td>0.77</td>
<td>0.63</td>
</tr>
<tr>
<td>fs_final_selection</td>
<td><strong>0.82</strong></td>
<td><strong>0.76</strong></td>
</tr>
</tbody>
</table>

**Discussion**

In this study, we used HRVTS data collected over 2 hours to improve the 30-day IHM prediction of septic patients on continuous ECG monitoring in the ED. An optimized feature set comprising 12 selected HRVTS features performed better than baseline feature sets which only included patient demographics, vital signs and single time-point HRV measures obtained at triage. This improved risk stratification could be used to inform management and disposition, such as early antibiotic therapy and ICU admission for identified high-risk patients.

While more research is needed to understand the physiological meaning and significance of each of the HRV parameters, studies have shown that HRV measures correlate with autonomic function and are predictive of adverse outcomes in septic patients. We hypothesize that higher-order dynamics and interactions in an HRVTS, including relative changes and trends over time may carry more prognostic information than the actual values of the HRV parameters themselves. This may explain why delta-based and regression-based features derived from HRVTS appear to be important risk predictors in our model.

We comprehensively explored various methods of feature engineering and selection to synthesize and make sense of the HRVTS data. In particular, our featurization method transforms a large number of data points for each patient into 15 key predictors and handles missing data. With the widespread adoption of wearable devices connected to the Internet of Things, we believe this type of high-resolution, high-volume time series data will become increasingly available. Therefore, the importance of our work lies in developing a methodology for processing and analysing HRVTS data which could be generalizable to other clinical settings and outcomes of interest, rather than in presenting a predictive model to be applied globally.

Our outcome of interest was IHM within 30 days during the same admission where the vitals and ECG were taken. Some studies did not specify a time period for mortality or if it was strictly within the same admission or not. We chose this endpoint as it is more likely to be sepsis-related compared to an out-of-hospital mortality or mortality from...
a subsequent admission. It is also more meaningful for physicians in terms of administering possible interventions such as closer monitoring and more aggressive management of high-risk patients.\\n
Our study had several limitations. Firstly, this was a single-institution study with a small sample size. Larger multi-centre prospective studies are required to validate our results. Secondly, we had included patients in our study based on clinical suspicion of sepsis and meeting at least 2 of the 4 SIRS criteria. Sepsis largely remains a clinical diagnosis and there is no gold standard to determine whether a patient is septic. Other studies have attempted to address this issue by including only patients with administered intravenous antibiotics, blood culture investigations or confirmed source of infection. We acknowledge that our cohort definition reflects suspected sepsis rather than confirmed sepsis. However, we believe this is reasonable given our eventual application in the ED where laboratory test results and confirmed diagnoses may not yet be available. In addition, while the SIRS criteria has recently been replaced with a new state of sepsis, defined as a life-threatening organ dysfunction caused by a dysregulated host response to infection, the usefulness of the SIRS criteria in diagnosis of sepsis was still emphasized by the same task force.

Thirdly, even though HRVTS can improve mortality prediction in suspected sepsis patients as shown in this study, they require more time and effort to obtain than single time-point HRV measures. Currently, we are developing a wearable device which can continuously monitor a patient’s ECG and automatically perform HRV analysis in real-time.

Fourthly, the time frame of 2 hours was chosen in this study based on the duration most patients spend in the ED while awaiting admission. Nevertheless, in our future studies, it is definitely worthwhile to explore the utility of HRVTS over a much longer period (such as 24-72 hours) and extending into the inpatient ward setting, as well as assess whether HRV-based risk predictions impact clinical management and outcomes. As continuous ECG data is inherently noisy (for example, due to patient movement or temporary disconnection), we chose to split the ECGs into 8 intervals of 15 minutes so that there is sufficient temporal window within each interval to capture a valid 5-minute segment for HRV analysis. However, this could have resulted in loss of information. Further study is needed to understand how we can better achieve granularity with noisy data.

Lastly, LASSO may arbitrarily select features from collinear groups, and further efforts should be put into interpreting the selected features and their odds ratios.

**Conclusion**

Serial HRV measurements over 2 hours can be used to improve prediction of 30-day IHM among septic patients in the ED compared to single time-point HRV measures taken at triage. This approach could be used in the ED to identify high-risk septic patients for appropriate management and disposition. In addition, the feature extraction and selection pipeline used in this paper could be applied to other short, noisy and incomplete time series data.

**Acknowledgments**

We would like to thank our research assistant Ms. Kavita Govinthurasah for her contribution in pre-processing the data for this project. This work is partially supported by National University of Singapore Start-up Grant WBS R-608-000-172-133.

**References**

Persuasive Data Videos: Investigating Persuasive Self-Tracking Feedback with Augmented Data Videos

Eun Kyoung Choe, Ph.D.1, Yumiko Sakamoto, Ph.D.2, Yanis Fatmi, B.S.3, Bongshin Lee, Ph.D.4, Christophe Hurter, Ph.D.5, Ashkan Haghsenas, B.S.2, Pourang Irani, Ph.D.2

1University of Maryland, College Park, Maryland, USA; 2University of Manitoba, Winnipeg, Manitoba, Canada; 3ENAC, University of Toulouse, France; 4Microsoft Research, Redmond, Washington, USA

Abstract

Self-tracking feedback with engaging and persuasive visualizations not only helps convey data but can also affect people’s attitudes and behaviors. We investigate persuasive self-tracking feedback by augmenting data videos (DVs)—novel, engaging storytelling media. We introduce a new class of DVs, called Persuasive Data Videos (PDVs), by incorporating four persuasive elements—primary task, dialogue, system credibility, and social supports—drawn from the Persuasive System Design Model. We describe the iterative design of PDVs and a within-subjects preliminary validation to check their persuasive potential. We then assess PDVs’ feasibility using the Persuasive Potential Questionnaire in a between-subjects study comparing a PDV against a conventional DV on Amazon Mechanical Turk (N = 252). Our results indicate the feasibility of using PDVs in providing individuals’ self-tracking feedback to convey persuasive health messages, based on which we discuss opportunities for designing persuasive behavioral feedback in an engaging way.

Introduction

Self-tracking technologies such as wearable devices and mobile health apps are advancing rapidly, enabling people to collect a wide range of personal data. Compared to a large amount of research and advancements in the data collection and sensing side of self-tracking technologies, we have not seen much innovation in designing effective self-tracking feedback: existing approaches in feedback designs are limited to simple counts, graphs, stylized representations, and texture feedback. Furthermore, current feedback designs portrayed in commercial trackers often lack persuasive elements delivered in an engaging way.

Designing engaging and persuasive self-tracking feedback is important in enhancing the efficacy of self-tracking. Engaging with personal data can provide personal insights and promote self-reflection. Moreover, well-designed feedback can help people alter their self-efficacy and achieve goals. In this work, we investigate ways to design engaging and persuasive self-tracking feedback.

Data videos (DVs) are novel, popular storytelling media, which would be beneficial in presenting self-tracking data. Being recognized as one of the seven genres of narrative visualizations, data videos combine both visual and audio stimuli to promote data-driven stories in an engaging manner. They also have the capacity to guide people’s cognitive processing for the data interpretation, and thus could make self-tracking data easily digestible.

While we employ data videos to deliver easy-to-understand self-tracking feedback in an engaging way, we specifically aim to realize their persuasive potential. In creating Persuasive Data Videos (PDVs), we augment data videos with persuasive elements drawn from the Persuasive System Design (PSD) model. This model was created in response to the upsurge in the development of persuasive software applications to aid in their design and evaluation. We carefully extracted four persuasive constituents from the PSD model—primary task support, dialogue support, system credibility support, and social support—that fit the context of delivering self-tracking data.

In this paper, we aimed to answer two research questions: (1) can data videos (DVs) convey persuasive elements drawn from the PSD model (i.e., the feasibility of creating PDVs), and (2) do persuasive data videos (PDVs) demonstrate higher persuasive potential than data videos? To answer these questions, we conducted a between-subjects study on Amazon Mechanical Turk (AMT) with 252 participants. We used the Persuasive Potential Questionnaire (PPQ) to assess DV’s and PDV’s persuasive potential and individual’s susceptibility to persuasion. Our study results indicate that participants in the PDV condition recognized the persuasive elements in the video at a higher level than those in the DV condition. Moreover, PDV demonstrated higher potential to persuade than DV.
The key contributions of this paper are twofold. First, we introduce PDVs, a new class of self-tracking feedback framework that includes four elements of persuasion, carefully adopted from the PSD model. Second, we provide the evaluation of PDVs with AMT participants, demonstrating their general persuasive potential in influencing people’s attitudes.

Related Work

**Self-Tracking Feedback**

Self-tracking (or self-monitoring) is an activity of recording one’s behaviors, thoughts, or feelings. Through self-tracking, people can enhance awareness, which may lead to self-discoveries and positive behavior changes. Providing effective feedback can augment such process, which is why designing effective self-tracking feedback is important for designers and researchers working in this field.

Many existing approaches in feedback design rely on text-based numerical representations, often augmented with simple visualizations. Some research on self-tracking feedback goes beyond simply presenting the collected data. For example, researchers inspired by the well-known “Framing Effects” investigated how different framing of self-tracking data (e.g., emphasizing either a positive or negative aspect of a target behavior) affects people’s perceptions and behaviors. They found that a feedback design with a particular framing could enhance self-efficacy or promote behavior change. These feedback helps people understand their current state, and furthermore, change behavior, but they are often static and lack actionable insights—a reason people may abandon self-tracking.

In this light, we aim to design persuasive and engaging self-tracking feedback. We explicitly incorporate persuasive elements and measure their persuasive potential, as we employ an engaging medium—data video—that could be appropriate to convey the summary of aggregated data.

**Persuasive System Design**

Persuasive technology concerns “computerized software or information systems designed to reinforce, change or shape attitudes or behaviors or both without using coercion or deception.” Theories and practical applications of persuasive technology have been explored in preceding decades. Such technologies span from social media apps such as Facebook that persuades its active user base to regularly upload pictures and share personal information to mobile apps designed to persuade people to lead more active lifestyles.

The Persuasive System Design (PSD) model expands the earlier work of Fogg and provides a framework to assist in the design of persuasive systems. The PSD model is composed of four persuasive categorical elements, namely (a) Primary Task Support, (b) Dialogue Support, (c) System Credibility Support, and (d) Social Support. Each of these categorical elements further contains psychological constructs known to enhance the persuasive potential of systems, which we adopt in the design of PDVs.

Also based on the PSD is the Persuasive Potential Questionnaire (PPQ), a subjective measurement tool that can aid in assessing the persuasive potential of a system. The PPQ is composed of fifteen questions grouped along three dimensions: (1) individuals’ susceptibility to persuasion which represents the degree by which an individual can be persuaded by a given feedback mechanism (IS); (2) the general persuasive potential of a system (GPP); and (3) the individual persuasive potential of the user (IPP). We build on the PSD to design our persuasive data videos and further assess their potential to persuade using the PPQ.

**Data Videos**

Data visualizations have been applied as a means to analyze and present personal data. Previous research suggests that data presentations tailored to an individual can facilitate motivation and possibly promote behavioral change. Data visualizations as a storytelling medium—data-driven storytelling or narrative visualization—have in recent years become increasingly popular among data scientists and journalists.

In particular, largely employing short animated data clips (or data-driven motion graphics), data videos appeal to a broad range of audiences as an emerging data-driven storytelling medium. When integrated with camera motion effects, such as zooming, audio stimuli, and a well designed narrative, data videos can effectively communicate data-driven facts in a short period of time. In addition, with appropriate attention cues, they can captivate and engage audiences.

We argue that data videos can further be equipped with the potential to persuade their viewers when conveying self-tracking data. The novelty of our work is in identifying those visual and narrative elements that can be integrated into data videos while offering self-tracking data feedback, and in evaluating their persuasive advantage.
Figure 1: The top row shows a regular Data Video (DV), while the bottom row shows its counterparts—the frames from a Persuasive Data Video (PDV). Each column of the bottom row represents one of the four persuasive elements we incorporated in the design of PDVs: (a) Primary Task Support; (b) Dialog Support; (c) Credibility Support; and (d) Social Support.

Persuasive Data Video

Persuasive Elements Selection

To design Persuasive Data Videos (PDVs), we carefully selected only those transferable persuasive constructs from the Persuasive System Design (PSD) framework. Transferable constructs refer to those that can be applied to a narrative structure and visual effects that make up a data video, without distracting from the core message, in our self-tracking context. Four major categorical elements from the PSD matched our criteria: (a) Primary Task Support; (b) Dialogue Support; (c) System Credibility Support, and (d) Social Support. Each of these four categories further contains specific persuasive constructs.

Below, we summarize these four categories and their constructs we applied to design persuasive data videos (Table 1).

From the Primary Task Support, we included the following constructs: Reduction, Personalization, and Simulation. These constructs are designed to encourage people in carrying out their task. For example, Figure 1-a demonstrates how we emphasized Personalization in the PDV by indicating “Your Wednesday class ends at 1:00pm. A perfect time for a walk!” with step counts visualization, whereas the DV’s message was simply set to “You are somewhat active,” while showing the same step count data.

From Dialogue Support, we included Praise, Rewards, Suggestion, Similarity, Liking, and Social Role. A system with these constructs can offer a dialogue to support a user progress toward their goals. For example, providing Praise via giving compliments such as “Great Job!” with a virtual trophy should lead people to a more positive attitude. We included messaging such as “Walk from home to the UofM. It’s 2.7 km!” as a means to support Suggestion (Figure 1-b).

From System Credibility Support, we included Trustworthiness, Expertise, Surface Credibility, Real-World Feel, Authority, Third-Party Endorsements, and Verifiability. Application of these constructs should endorse overall credibility of the system, and thus enhance persuasiveness. For example, to emphasize Surface Credibility, the narrator in our PDV was Dr. McKenzy, a hypothetical physician, who carries the persona of someone that people could trust to provide health recommendations (Figure 1-c). We also considered other representations of authority, such as to suggest the information coming from the World Health Organization or other well-known and credible sources. However, their identity was difficult to convey with a cartoon-style character, and stronger impressions could be delivered through inclusion of world-renowned experts.

Finally, from Social Support, we included Social Learning, Social Comparison, Normative Influence, Social Facilitation, Cooperation, Competition, and Recognition. Applying these constructs should motivate people via social influences. For example, in our PDV, Dr. McKenzy informs the viewer that they have 839 steps to beat Lisa (Figure 1-d). This instills a degree of Social Comparison by letting them observe their accomplishments in light of those of others.

We settled on these mappings through iterative discussion among authors, and by experimenting and considering the various visual options available in Animaker, the animated video authoring tool we used. In the next section, we describe how we realized the PDV with the self-tracking data through a preliminary validation.
Table 1: Four major elements and 23 constructs from the Persuasive System Design framework. *Similarity was removed in the revised Persuasive Data Video.

<table>
<thead>
<tr>
<th>Elements</th>
<th>Constructs</th>
<th>Descriptions of how the constructs were applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Task Support</td>
<td>Reduction</td>
<td>Provided activity with instructions.</td>
</tr>
<tr>
<td></td>
<td>Personalization</td>
<td>The persona in the video talked specifically to the person.</td>
</tr>
<tr>
<td></td>
<td>Simulation</td>
<td>Provided clear tips which should lead to the desired outcome.</td>
</tr>
<tr>
<td>Dialogue Support</td>
<td>Praise</td>
<td>Provided praise when the person's behavior was positive.</td>
</tr>
<tr>
<td></td>
<td>Rewards</td>
<td>Provided virtual prizes when the person’s behavior was positive.</td>
</tr>
<tr>
<td></td>
<td>Suggestion</td>
<td>Suggested activities which should lead to the desired goal.</td>
</tr>
<tr>
<td></td>
<td>Similarity*</td>
<td>Used relatively casual language.</td>
</tr>
<tr>
<td></td>
<td>Liking</td>
<td>The video was designed to be visually attractive with data visualization effects.</td>
</tr>
<tr>
<td></td>
<td>Social role</td>
<td>The persona in the video was a doctor or professor.</td>
</tr>
<tr>
<td>System Credibility</td>
<td>Trustworthiness</td>
<td>Provided data from the person’s weekly activity records.</td>
</tr>
<tr>
<td></td>
<td>Expertise</td>
<td>Provided quotes and suggestions taken from research.</td>
</tr>
<tr>
<td></td>
<td>Surface credibility</td>
<td>The video was designed to look competent.</td>
</tr>
<tr>
<td></td>
<td>Real-world feel</td>
<td>Provided actual organization and authors.</td>
</tr>
<tr>
<td></td>
<td>Authority</td>
<td>Provided references to scientific studies.</td>
</tr>
<tr>
<td></td>
<td>Third-partyendorsements</td>
<td>Provided well-known organization names.</td>
</tr>
<tr>
<td></td>
<td>Verifiability</td>
<td>Provided organization names with links.</td>
</tr>
<tr>
<td>Social Support</td>
<td>Social learning</td>
<td>Provided friends’ ranking results.</td>
</tr>
<tr>
<td></td>
<td>Social comparison</td>
<td>Provided graphs that presented the person’s data against their friends’ data.</td>
</tr>
<tr>
<td></td>
<td>Normative influence</td>
<td>Provided relevant health information to the person and their friends.</td>
</tr>
<tr>
<td></td>
<td>Social facilitation</td>
<td>The person’s friends performance were made salient.</td>
</tr>
<tr>
<td></td>
<td>Cooperation</td>
<td>The person participated in a program with their friends.</td>
</tr>
<tr>
<td></td>
<td>Competition</td>
<td>The person’s performance was compared against their friends.</td>
</tr>
<tr>
<td></td>
<td>Recognition</td>
<td>Each person’s accomplishment was celebrated in the video.</td>
</tr>
</tbody>
</table>

**Persuasive Data Video Development**

We developed Persuasive Data Videos (PDVs) to examine whether the PSD elements described above can be properly implemented, and thus be perceived by the viewers. To ensure that the persuasive impact is the result of the persuasive elements we employed, we compared the PDV against the regular data video (DV), which was bereft of any persuasive elements. We developed the video stimuli to suit an online experiment environment (i.e., Amazon Mechanical Turk).

**Preparing the Videos.** Both the PDV and DV presented the summary of an imaginary person’s health data (i.e., Cameron, 22 years old). While the DV primarily focused on the presentation of weekly health summary data (step count, heart rate, & sleep duration), the PDV further included persuasive elements (as we detailed in the earlier section). The video structure had the following sequence of main items: An introduction showed the average number of steps taken during the week using a line graph and a bar chart. This was followed by a bar chart displaying the average number of hours slept. A line graph was then used to display the daily average heart rate during the week. And finally, a data summary of week 1 was presented and compared to recommendations (PDV: tailored recommendation, DV: generic recommendation).

We also ensured that both videos had the same length (1 min 59 sec), with the same background music, displayed data, and structure. Both videos were constructed with the same self-tracking data provided by one of the authors. The videos were created using Animaker, an online tool that allowed us to pick characters, narration styles, and animated data charts.

**Preliminary Validation.** We assessed the implementation of persuasiveness with seven HCI graduate students (all males; 22 to 41 years old), who were unfamiliar with the goal of the study and the PSD. To conduct the assessment, we generated 46 statements, two for each of the 23 constructs of the four PSD persuasive elements (Table 1) to which participants responded using a 7-point Likert scale (1 = Strongly Disagree; 7 = Strongly Agree). For example, for the Liking construct, the two statements were “The video was visually attractive.” and “Overall, the video’s look and feel have appealed to you.”
Figure 2: Subtracted perceived persuasiveness (i.e., PDV - DV) from the preliminary validation. Positive numbers indicate higher persuasive elements in PDV.

We conducted this assessment using Qualtrics. Participants first read the following preamble:

“Imagine students at the University of Monty are participating in a health program, where their health related data are tracked using their smartwatches. At the end of each week, all the students receive summaries of their data in a video format. We would like you to imagine that you are watching the weekly review of a 22 year-old student, Cameron, after his first week. Cameron is participating in the program with a few of his friends, along with many other students.”

Participants watched the PDV first and then the conventional DV, to prevent possible ceiling effect. They watched full HD (1080p) videos uploaded to YouTube. Participants were allowed to replay the video after they finished watching the entire video once, but not allowed to skip or go back. After watching a PDV, they answered the 46 questions we prepared. Subsequent to this, they watched a DV, followed by the same set of 46 question items. At the end of the session, we asked for suggestions to improve the PDV with an open ended question.

Revising the PDV: The two question items within each construct were highly correlated (Cronbach’s $\alpha > .85$). Participants showed an equal liking to both video styles, and yet, persuasive elements were successfully recognized at a higher level in the PDV than in the DV, except for the Primary Task Support (Figure 2). This result affirms that our choice of PSD elements, and further, our interpretation of these into a data video medium has the potential to persuade. However, based on the preliminary assessment results and qualitative feedback, we made three changes to the stimuli for a large-scale study (see the next section).

First, we revised PDVs not to include the Similarity construct as it involves “imitating users in some specific way,” for example, by using phrases or slang used by the user, from everyday conversation. We realized that this was not possible as we did not personalize the PDV to each viewer; and furthermore, Similarity clashed with our Trustworthiness, Credibility, and Authority representation in the PDV condition. We wished to maintain a similar degree of persuasion without losing generality for the participants in the online experiment setting. Thus, we embedded 22 persuasive aspects in the revised PDV and not 23 as used earlier. Second, we removed the heart rate data to shorten the video length to 1 min 38 sec, thus leaving the video with only step count and sleep duration data. Third, we changed the character of the persona in the DV as the viewers mistook him (a Caucasian male wearing a bandana) as a fitness coach and thus unintentionally attributed him a higher credibility. The new character was also a Caucasian male but physically larger than the previous one.
Table 2: Adjusted Persuasive Potential Questionnaire.

<table>
<thead>
<tr>
<th>Susceptibility to Persuasion (SP)</th>
<th>Q1</th>
<th>When I hear others talking about something, I often re-evaluate my attitude toward it.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q2</td>
<td>I do not like to be influenced by others.</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>Persuading me is hard even for my close friends.</td>
</tr>
<tr>
<td></td>
<td>Q4</td>
<td>When I am determined, no one can tell me what to do.</td>
</tr>
<tr>
<td>General Persuasive Potential (GPP)</td>
<td>Q5</td>
<td>I feel that the video would make its viewer change their behaviors.</td>
</tr>
<tr>
<td></td>
<td>Q6</td>
<td>the video has the potential to influence its viewer.</td>
</tr>
<tr>
<td></td>
<td>Q7</td>
<td>the video gives viewer a new behavioral guideline.</td>
</tr>
<tr>
<td>Individual Persuasive Potential of the System (IPP)</td>
<td>Q8</td>
<td>This program is exactly what I need to change my attitude.</td>
</tr>
<tr>
<td></td>
<td>Q9</td>
<td>Thanks to the program I reach my goals.</td>
</tr>
<tr>
<td></td>
<td>Q10</td>
<td>I will use this program as often as I need to regulate my activity.</td>
</tr>
<tr>
<td></td>
<td>Q11</td>
<td>I think that I will also use such a program in the future.</td>
</tr>
<tr>
<td></td>
<td>Q12</td>
<td>I will use this program on a weekly basis.</td>
</tr>
<tr>
<td></td>
<td>Q13</td>
<td>This video does not cause a change in behavior with me.</td>
</tr>
<tr>
<td></td>
<td>Q14</td>
<td>This video causes me to do certain things differently.</td>
</tr>
<tr>
<td></td>
<td>Q15</td>
<td>With the help of the video, I will behave differently the next week.</td>
</tr>
</tbody>
</table>

Study

To investigate whether Persuasive Data Videos have higher potential to alter lay people’s attitude than conventional data videos, we conducted a between-subjects study via Amazon Mechanical Turk (AMT). We were primarily interested in their “potential” only, as a longitudinal study would be necessary to assess their effective persuasive power over an extended time.

Procedure and Participants

After the participants read the consent form and agreed to participate, they read the following preamble:

“Imagine some people are participating in a health promotion program, where their health-related data are tracked using their smartwatches. At the end of each week, they receive summaries of their data in a video format. We would like you to imagine that you are watching the weekly review of a 22-year-old Alex, after his/her first week. Alex is participating in the health promotion program with a few of his/her friends, along with many other people.”

Subsequent to this, they watched either a PDV or a DV, and answered questions about the video and their demography. We ensured that participants could only view either a PDV or DV.

We recruited AMT workers who resided within either Canada or the US with HITs (or Human Intelligence Tasks) approval rate of 97% or above, and speak English. Participants recruited for both conditions saw the identical study description: “First, you will be asked to watch a short Video Clip (approximately 2 minutes). Subsequently, you will be asked to complete a survey asking you about your background, and experience of watching the video clip.” Altogether, 339 AMT Workers participated in this study. Following the recommendation to enhance response quality on AMT through inclusion of verifiable questions, we excluded AMT workers who did not pay careful attention to the study instructions as indicated by two gotcha questions (e.g., “How many words are in this sentence?”). Out of 339, 87 participants did not pass our two gotcha questions. This left us with 252 participants (PDV = 123; DV = 129), and their age ranged between 22 and 60 (PDV: \( M = 35.07, SD = 9.37 \); DV: \( M = 36.26, SD = 9.68 \)). In both conditions, 42% of participants were female. Participants who completed the study received 1.54 USD as a compensation, and they spent 8 mins 44 secs on average.

Questionnaire

We adapted the Persuasive Potential Questionnaire (PPQ) to match the presentation and assessment of PDVs, resulting in the 15 questions (Table 2). This allowed us to measure (1) individuals’ susceptibility to persuasion (SP: four questions), (2) the general persuasive potential of the system (GPP: three questions), and (3) the persuasive potential of the system to the individual (IPP: eight questions), using a 7-point Likert scale (1 = Strongly Disagree; 7 = Strongly Agree). We added one question to ask participants’ interest in improving their own health, and worded it as follows: “I am interested in improving my health” again on a 7-point Likert scale.
Subtracted perceived persuasiveness (i.e., PDV - DV) from the AMT study. Positive numbers indicate higher persuasive elements in PDV.

Next, to investigate whether the persuasive elements were embedded appropriately, we asked the same questions used in our prior assessment about the persuasiveness of the video. Note, two questions were used to investigate one construct in the prior validation where we found a high inter-item correlation within each construct ($r_s > .85$). However, since we added 15 PPQ question items in this study, we were concerned about participants’ potential cognitive fatigue and boredom, especially because we conducted this study using an online format. Hence, we removed the overlapping corresponding question items and made minor adjustments (rephrasing the combined question only slightly).

**Results**

First, we investigated the feasibility of creating PDVs by embedding persuasive elements (i.e., can data videos (DVs) convey persuasive elements drawn from the PSD model). Since the data was not normally distributed, we conducted non-parametric analyses. Results from Mann-Whitney U tests confirmed that participants in the PDV condition perceived all the persuasive elements at a higher level than those who were in the DV condition ($p < .05$; Figure 3).

Next, we investigated whether participants in the PDV condition perceived higher persuasive potential for attitude change than those who were in the DV condition. First, we investigated Cronbach’s alphas for all three aspects: (1) individual’s susceptibility to persuasion (SP) ($\alpha = .75$); (2) the general persuasive potential (GPP) of the system ($\alpha = .86$); and (3) the persuasive potential of the system to the individual (IPP) ($\alpha = .84$). Since all three scales were acceptable or good, we created aggregates for each of the three aspects. The results from Mann-Whitney U tests (Figure 4) showed that the video type did not have any effect on the individual’s susceptibility to persuasion ($p = .47$), indicating that participants in the PDV condition and DV condition were equally susceptible to persuasion. Next, we found that for participants’ general persuasive potential (GPP) of the system, those in the PDV condition indicated higher perceived general persuasion ($Mdn = 5.67$, $n = 123$) relative to its counterpart ($Mdn = 5.33$, $n = 129$), $U = 7520$, $z = -2.83$, $r = .18$. Finally, no significant video type effect was found for persuasive potential of the system to the individual (IPP) ($p = .30$).

**Discussion and Future Work**

**Reflecting on Study Results**

After making the necessary adjustments to our Persuasive Data Video (PDV), the manipulation check (i.e., “a measure used to determine whether the manipulation of the independent variable has had its intended effect on a participant”) indicated that the persuasive elements had the effect we expected. Furthermore, participants who watched a PDV in the AMT study perceived its higher potential to influence the general population, compared to those who watched a Data Video (DV). Altogether, we
believe that our PDV might offer beneficial feedback for applications involving attitudinal change.

Although PDVs’ potential to influence the general population (i.e., others) was perceived as higher than that of DVs, this video type difference was not found when it concerns the video’s potential to persuade oneself. This inconsistency between perceptions regarding self vs. others could be attributed to the third person effect whereby individuals normally perceive that others are more readily influenced by media than themselves.[12, 13] According to these prior works, individuals estimate being shielded from such effects but not their peers. We intend on following up with this result as it may be possible that the PDV had an unmeasured effect on its viewers. Altogether, while assessing behavior change is outside the scope of this investigation, we believe that PDVs have the potential to influence peoples’ attitude.

Overall, we attribute our outcomes to the use of the Persuasive System Design (PSD) framework. Specifically, the framework that PSD provides may have been particularly suitable for incorporating in data videos, as we did here. Unlike conventional feedback visualizations, videos allow us to spread the persuasive elements gradually, and throughout the entire viewing period. Videos also implicitly guide the viewer to those critical aspects of the data and as such can carefully control the amount of information delivered to the user. Further work is needed to examine which specific PSD elements and their video representation had higher potential to persuade.

Persuasive Feedback for Self-Tracking Data

A key strength of our approach is the use of data videos, a medium that resonate well with younger generations (including kids) who routinely consume videos. Data videos could also be delivered effectively on mobile phones, where people commonly track and view their self-tracking data. On the other hand, data videos are not suitable for delivering feedback in real-time, and thus, should be used for aggregated data that fosters reflection on past behaviors. Also, even with this engaging medium, people might get bored over time and disengage. To maintain viewers’ interest, we need to identify a desirable viewing cycle (e.g., once a day, a week, or a month) that can reveal interesting and surprising aspects about oneself, and which may depend on the data type.

In addition, even though there have been efforts on easing the process of creating data videos, it is not easy to create data videos especially with personalized recommendations. An important direction for future research would be to develop a PDV authoring tool to cover diverse contexts (e.g., productivity, finance, screen time). Such video authoring tools will enable researchers to conveniently produce personalized PDVs, which would be beneficial for conducting deployment studies.

Limitation and Next Step

Although our overall results are encouraging, we acknowledge that this study is only an initial step. Contextual limitation of this work is using the hypothetical scenario and measuring attitude rather than actual behavior. Our scenario conveyed the data of a 22 years old self-tracker, which may not be representative of AMT workers. However, because we are still at the early stage of designing effective PDV, we argue that conducting a field deployment study to measure behavioral outcomes is not the best first approach. Moreover, as demonstrated in prior works, hypothetical scenarios have been used to test a variety of feedback design contestants to identify the most effective one before running a deployment study.[11]

In this work, conducting an AMT study using a hypothetical scenario allowed us to recruit a large number of participants with relatively low cost and helped us understand the effect of persuasive elements in a brief period. Our study will help designers and researchers create engaging, persuasive data videos, which could be embedded in a self-tracking technology for a long-term
deployment study. We intend to enhance our PDVs and examine their ability to influence behavior, and plan to run a longitudinal study, for which PDVs will be tailored to each participant’s data and context. Such investigations will further advance our knowledge about new possibilities with PDVs, and will particularly be beneficial to designers of self-tracking applications.

Conclusion
We introduced Persuasive Data Video (PDV), an enhanced form of Data Video (DV) augmented with persuasive elements, for self-tracking feedback. Results from our M-Turk study were comparatively favorable toward PDV, at least indicating the potential of PDV in influencing people’s attitude. We believe that these benefits were found because of two main reasons. First, we successfully translated persuasive elements from the Persuasive System Design into our PDV. Second, related to our first point, videos may be a particularly suitable format for applying the PSD framework. Our results motivate several directions for future work in the context of designing effective self-tracking feedback, including a consideration of effective viewing cycle of PDVs, designing a PDV authoring tool that supports convenient production of personalized PDVs, as well as further studies to assess the efficacy of PDVs on people’s behavior change in a real-world context.

References
17. T. Dhawan, B. G. Choi, and R. J. Katz. Diet and nutrition smartphone apps demonstrate poor adherence to the 2015-2020
dietary guidelines for americans, 2016.
35. Qualtrics. The leading research amp; experience software — qualtrics, 2018.
Engaging heart failure patients from a clinical data research network: A survey on willingness to participate in different types of research

Yong K. Choi, PhD MPH¹, Javier E. López, MD MAS¹,², Daniella Meeker, PhD³, Lucila Ohno-Machado, MD PhD⁴, Katherine K. Kim, PhD MBA¹, and the pSCANNER Team
¹University of California Davis, Sacramento, CA; ²Division of Cardiovascular Medicine, University of California Davis, Sacramento, CA; ³University of Southern California, Los Angeles, CA; ⁴University of California San Diego, La Jolla, CA

Abstract

The willing participation of patients in clinical research is a critical element in national efforts to collect health data for precision medicine and large cohort studies. However, recruiting patients is challenging. Clinical data research networks (CDRN) have primarily been used for observational studies, but may be able to enhance recruitment efforts. We need a better understanding of patient motivation and preferences for research participation and their interest in different types of research activities, particularly among those who are already represented in CDRNs. We surveyed a heart failure patient cohort constructed from EHRs in a CDRN to assess research participation. Results showed that CDRN recruitment is feasible. Respondents were most interested in completing a one-time survey and giving a blood sample one time. They were least interested in a study about weight control that require surgery. We found statistically significant associations between race and research activity interests.

Introduction

The goal of precision medicine research is to better understand the impact of individual variability in genes, environment, and lifestyle on disease prevention and treatment. The success of this type of research is heavily dependent upon the willingness of large numbers of volunteers to join research initiatives and share their genetic, health and lifestyle data. Electronic health records (EHR) from multiple health care centers, when aggregated under a standard data model, can facilitate identification of large-scale cohorts. The establishment of clinical data research networks (CDRNs) offers an opportunity to leverage EHR resources and lower the technical and regulatory barriers for research. However, first generation CDRNs such as the Health Maintenance Organization Research Network (HMORN) Virtual Data Warehouse¹ and the U.S. Food and Drug Administration (FDA) Sentinel Initiative research network² have primarily utilized retrospective observational data from existing clinical data and did not engage participants directly.

Enrolling patients into research studies has traditionally been challenging. In fact, Williams et al. reported that insufficient accrual rate was one of the major reasons for terminated clinical trials reported in ClinicalTrials.gov database.³ In other types of research, such as studies based on surveys, which require less effort than clinical trials on the part of participants, response rates are quite low, hovering around 9% for telephone surveys.⁴ A number of studies have shown that racial and ethnic minorities in the US are less likely to express willingness to participate in a health-related studies.⁵⁻⁶ Underrepresentation of minority groups has also been identified in previous clinical trials.⁷⁻⁸ Several studies have shown that people with higher education level were more willing to participate in medical research.⁹⁻¹⁰ In addition, participants were more likely to share data via a research network if requests came from hospitals, universities, or medical groups, as they had trust in such organizations.¹¹ Thus, one concern is whether CDRNs can effectively engage diverse patients and enroll them in research studies.

The purpose of this study was to examine if a cohort of patients with heart failure identified from the EHRs of a large CDRN could be engaged to participate in new research, and to assess what type of research activities they preferred. We surveyed patients from the patient-centered SCAlable National Network for Effectiveness Research (pSCANNER) on their willingness to participate in future research. pSCANNER is a stakeholder-governed, privacy-preserving, distributed CDRN with access to EHRs across institutions and on over 30 million patients from all 50 states¹²⁻¹³ that was initially supported by the PCORNet.¹⁴ Our rationale for surveying participants was that understanding challenges from their perspectives could help CDRNs improve recruitment and retention for future studies.
Methods

In the pSCANNER network, heart failure was a condition of interest to investigators, and several future studies in this condition were planned. We first identified a cohort of patients with heart failure by running a query written against the Observational Medical Outcomes Partnership Common Data Model (OMOP) using ICD-9/ICD-10 code and the following inclusion criteria: (1) Age ≥ 18 years old; (2) Two reported diagnoses (inpatient or outpatient) related to HF (ICD-9 CM codes 398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 414.8, 428.x.) between 2011 and 2016; OR one principal hospital discharge diagnosis of heart failure between 2011 to 2016, for patients not known to be deceased.

The identified participants were recruited first via e-mail (if the email address was available), using a consistent template identifying the study as conducted by pSCANNER and signed by a local investigator. A telephone outreach firm conducted phone recruitment and completion of the survey for patients for whom a valid email address was not available, or who did not respond to an email. A total of five attempts were made to contact participants. IRB approval was obtained at each of the pSCANNER sites that participated in the study: five University of California health systems (UC San Diego, UC Davis, UC Los Angeles, UC Irvine, and UC San Francisco), Veterans Administration Health System in San Diego, and the University of Southern California, Keck School of Medicine including AltaMed, a USC-partnered community clinic. For one of the sites, only in-person recruitment with paper and pencil survey collection was conducted.

The survey was developed by the PCORnet Survey Team (KK was a member), a collaborative group of investigators from multiple CDRNs participating in PCORnet. Items on the survey were intended for use by cohorts of people who were overweight/obese and by those with other common chronic conditions, such as cardiovascular disease. The items included: (1) key demographics and self-reported health status, (2) participants’ interest to participate in a variety of research activities (19 activities rated using a 3-point Likert response scale from 1=not interested to 3=very interested), and (3) preferred mode of contact for future research with yes/no response categories. The order in which each of the research activities were presented was randomized to reduce bias. The survey took approximately 5 to 7 minutes to complete. The survey was implemented in a centralized Qualtrics server and de-identified, individualized links created for each participant candidate. Participants who accessed the survey website could review the purpose of the project, the elements of consent, and the process for withdrawing from the cohort and for completing the survey. The survey responses collected by the phone interviewer were entered into Qualtrics using the unique, de-identified link. There was no time limit to the survey. All survey data were collected and stored in standardized formats under rigorous security protocols. Survey responses, including self-reported identifiers, were retained by each site’s research personnel on a password-protected secure server. Non-identifiable survey responses were attached to the pSCANNER database for individuals who responded. Direct identifiers reported on the survey (email address and phone number) were not attached to the electronic record.

All statistical analyses of data were performed in SPSS (SPSS, Chicago, IL, USA). The socio-demographics and health characteristics and mode of contact questions were summarized as numbers and percentages. Two-way analysis of variance (ANOVA) with post-hoc Tukey honest significant difference (HSD) was performed to compare the mean total scores of Research Activity Interest questions and to check for a difference between groups for two categorical demographic covariates: race and education. Statistical significance was accepted as p<0.05.

Results

Completion rates for the survey are noted in Table 1. The survey was completed by 2,444 respondents (10.2%) of the total 23,955 heart failure cohort. Of the 2,444 people who completed the survey, 68 were aged under 18 and did not meet the eligibility criteria. Therefore their survey data were excluded from further analysis, leaving 2,376 completed surveys. Table 1 shows response rates by each of the 7 pSCANNER organizations.

<table>
<thead>
<tr>
<th>Table 1: Survey cohort breakdown by recruiting organization (A-G)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heart Failure Cohort Survey</strong></td>
</tr>
<tr>
<td>HF Total</td>
</tr>
<tr>
<td># Surveys Completed</td>
</tr>
<tr>
<td>% Completed</td>
</tr>
</tbody>
</table>

306
Table 2 summarizes the socio-demographic and health characteristics of survey respondents (n=2,376). The mean age was 63 years (SD=15.2, range 18-97) and majority were male (59%). The respondents were likely to be Caucasian (58%) and college educated (69%). About half of respondents (44%) were married at the time of completing the survey. Most of the respondents (60%) were overweight or obese (having BMI $\geq 25$) and 867 (36.5%) were diagnosed with diabetes. Among the 2,376 respondents, 756 (32.2%) indicated that they had previously participated in research.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in Years, $M$, (SD), range</td>
<td>2,376</td>
<td></td>
</tr>
<tr>
<td>63.0, (15.2), 18-97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>963</td>
<td>40.5</td>
</tr>
<tr>
<td>Male</td>
<td>1,395</td>
<td>58.7</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>Ethnicity and Race (Self-Report)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1,376</td>
<td>57.9</td>
</tr>
<tr>
<td>Hispanic</td>
<td>447</td>
<td>18.8</td>
</tr>
<tr>
<td>Black</td>
<td>294</td>
<td>12.4</td>
</tr>
<tr>
<td>Asian</td>
<td>104</td>
<td>4.4</td>
</tr>
<tr>
<td>Other</td>
<td>64</td>
<td>2.7</td>
</tr>
<tr>
<td>Native American</td>
<td>28</td>
<td>1.2</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>23</td>
<td>1.0</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>4</td>
<td>0.2</td>
</tr>
<tr>
<td>Prefer not to answer</td>
<td>23</td>
<td>1.0</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8th grade or less</td>
<td>134</td>
<td>5.6</td>
</tr>
<tr>
<td>Some high school</td>
<td>138</td>
<td>5.8</td>
</tr>
<tr>
<td>High school graduate or GED</td>
<td>444</td>
<td>18.7</td>
</tr>
<tr>
<td>Some college or 2-year degree</td>
<td>843</td>
<td>35.5</td>
</tr>
<tr>
<td>4-year college graduate</td>
<td>391</td>
<td>16.5</td>
</tr>
<tr>
<td>More than 4-year college degree</td>
<td>400</td>
<td>16.8</td>
</tr>
<tr>
<td>Currently Married</td>
<td>1,048</td>
<td>44.1</td>
</tr>
<tr>
<td>Physical Activity Level*, M (SD)</td>
<td>2,413</td>
<td></td>
</tr>
<tr>
<td>3.5 (1.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obese (BMI $\geq 30$)</td>
<td>917</td>
<td>33.4</td>
</tr>
<tr>
<td>Overweight (BMI 25 to &lt;30)</td>
<td>653</td>
<td>27.5</td>
</tr>
<tr>
<td>Diabetes Diagnosis (Yes/No)</td>
<td>867</td>
<td>36.5</td>
</tr>
<tr>
<td>Previous participation in research</td>
<td>756</td>
<td>32.2</td>
</tr>
</tbody>
</table>

*Rated from 1 = very inactive, 2 = a couple times a month, 3 = active most weeks, 4 = several days a week, 5 = most days
Table 3 shows participants’ interest in future research and their interests in different research activities, from 1= not interested to 3=very interested. Among the participants, the top five research activity interests were (1) completing a one-time survey or list of questions (very interested= 48.6%), (2) giving a blood sample one time (45.1%), (3) giving a blood sample that is used to study your DNA (43.6%), (4) testing a treatment given by phone or over the Internet (43.5%) and (5) completing a survey two or more times (40.4%). Participants were least interested in participating in a study about weight control that required surgery (not interested= 69.2%), a study that uses a medicine to help control weight (44.6%), and a project that involved other members in the family (42.3%).

The mean and standard deviation was calculated for each research activity to conduct a two-way ANOVA to examine the effect of race and education level on interest in research activities. There was a statistically significant association between the effects of participant race and the mean score for interest in all research activities ($F(8,2134) = 4.70, p < 0.0001$). However, no significant association was found between education and the total mean score ($F(6,2134) = 0.63, p = 0.71$).

### Table 3: Willingness to Participate in Research

<table>
<thead>
<tr>
<th>Interest in Research Activity* (n\textsuperscript{b})</th>
<th>Not interested, %</th>
<th>Somewhat interested, %</th>
<th>Very interested, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completing a one-time survey or list of questions (n=2,369)</td>
<td>5.5</td>
<td>45.9</td>
<td>48.6</td>
</tr>
<tr>
<td>Giving a blood sample one time (n=2,370)</td>
<td>12.1</td>
<td>42.8</td>
<td>45.1</td>
</tr>
<tr>
<td>Giving a blood sample that is used to study your DNA (n=2,353)</td>
<td>16.0</td>
<td>40.5</td>
<td>43.6</td>
</tr>
<tr>
<td>Testing a treatment given by phone or over the Internet i.e. like getting advice about your health (n=2,368)</td>
<td>14.6</td>
<td>41.9</td>
<td>43.5</td>
</tr>
<tr>
<td>Completing a survey two or more times (n=2,364)</td>
<td>12.9</td>
<td>46.7</td>
<td>40.4</td>
</tr>
<tr>
<td>A study about weight control that focuses on working on your diet or how active you are (n=2,317)</td>
<td>24.3</td>
<td>35.5</td>
<td>40.2</td>
</tr>
<tr>
<td>Taking part in a project that requires you to wear a device or monitor that collects information about your activities (n=2,363)</td>
<td>16.7</td>
<td>44.9</td>
<td>38.4</td>
</tr>
<tr>
<td>A study about weight control that tried to understand the genetics of obesity, and would require a blood sample from you (n=2,310)</td>
<td>31.4</td>
<td>31.0</td>
<td>37.6</td>
</tr>
<tr>
<td>Completing a weight related survey two or more times (n=2,306)</td>
<td>22.9</td>
<td>42.9</td>
<td>34.2</td>
</tr>
<tr>
<td>Testing a treatment where you need to come to clinic one or more times (n=2,365)</td>
<td>20.9</td>
<td>46.9</td>
<td>32.1</td>
</tr>
<tr>
<td>Taking part in a project where you have to respond to phone calls or text messages to provide information about what you are doing with diet or exercise or other issues about your health every day (n=2,370)</td>
<td>25.1</td>
<td>44.2</td>
<td>30.6</td>
</tr>
<tr>
<td>A study that uses medicines to help control weight (n=2,311)</td>
<td>44.6</td>
<td>28.6</td>
<td>26.9</td>
</tr>
<tr>
<td>Taking part in a project that involves meeting at a local community center or school (n=2,367)</td>
<td>28.9</td>
<td>45.2</td>
<td>25.9</td>
</tr>
<tr>
<td>Testing a treatment where you have to take a medicine or other treatment, come for clinic visits, AND give blood samples (n=2,366)</td>
<td>35.8</td>
<td>40.7</td>
<td>23.5</td>
</tr>
<tr>
<td>Testing a treatment where you have to take a medicine or other treatment, and come for clinic visits (n=2,366)</td>
<td>33.9</td>
<td>43.0</td>
<td>23.2</td>
</tr>
</tbody>
</table>
Taking part in a project that involves you and other people in your family (n=2,363) 42.3 37.1 20.6
Taking part in a project in which you would stay in the hospital for 1 or more days (n=2,366) 41.9 37.8 20.2
Taking part in a project which involves a procedure such as a special x-ray or new type of surgery (n=2,362) 39.9 42.0 18.1
A study about weight control that required surgery (n=2,292) 69.2 18.8 12.1

a Each activity was rated from 1 = not interested, 2 = somewhat interested, 3 = very interested
b Total number of people who responded to the survey item

Participants indicated the preferred modes of contact to learn about potential research studies (Table 4). Half of respondents (50%) preferred to be contacted by personal phone call from research staff or doctor. Patients were also willing to be contacted by email (35%), and letter or postcard in the mail (29.9%). Less than 2% of respondents did not wish to be contacted for future studies. The least preferred mode of contact was through social media (0.8%).

Table 4: Mode of Contact (n=2,376)

<table>
<thead>
<tr>
<th>Interest in Research Contact, by Modea</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal phone call from research staff or my doctor.</td>
<td>1,192</td>
<td>50.2</td>
</tr>
<tr>
<td>E-mail</td>
<td>821</td>
<td>34.6</td>
</tr>
<tr>
<td>Letter or post card in the mail</td>
<td>678</td>
<td>28.5</td>
</tr>
<tr>
<td>Cell phone text messaging</td>
<td>359</td>
<td>15.1</td>
</tr>
<tr>
<td>Talking face-to-face with research staff or my doctor when I am visiting the clinic</td>
<td>345</td>
<td>14.5</td>
</tr>
<tr>
<td>Other</td>
<td>147</td>
<td>6.2</td>
</tr>
<tr>
<td>A computer-created phone message</td>
<td>99</td>
<td>4.2</td>
</tr>
<tr>
<td>Do not contact me</td>
<td>42</td>
<td>1.8</td>
</tr>
<tr>
<td>Social media (such as Facebook, Twitter, or Pinterest)</td>
<td>20</td>
<td>0.8</td>
</tr>
</tbody>
</table>

a Percentages do not add up to 100% because multiple responses were allowed. 3,703

Discussion and Lessons Learned

The design and administration of our survey demonstrated the feasibility of identifying a clinical cohort from EHRs in a large CDRN and engaging them to participate in a new research topic.

Lessons Learned for the CDRN. While the IRB protocol was identical at the sites except for site G, where only in-person survey completion was approved, there were variations in implementation that may account for the differences in the response rate. At site G, the number of completed responses was the lowest due to the need for in-person staffing to recruit respondents during clinic hours. However, the in-person engagement appears to have resulted in very high response rate. Additionally, the overall response rate of 10.2% was comparable to the other surveys of CDRN cohort that incorporated broad recruitment approaches such as unsolicited e-mails to eligible patient population.\textsuperscript{15} Site A was the first site to field the survey. At this site, all patients were emailed at the same time with the letter signed by the pSCANNER PI and a physician, but not one who was associated with Cardiology. A number of calls were received from patients questioning why they had been contacted and whether the study was a legitimate research project from that institution. Based on the experience of site A, site C revised its invitation email to be co-signed by a cardiologist and sent out emails in four batches one week apart to allow for any
adjustments needed. After the first batch, cardiology clinics at site C also received a few calls and questions during clinic visits asking about the legitimacy of the study. Hence, site C developed an informational brief about the study and distributed it to all clinic staff and physicians so that they could respond knowledgeably to any patient questions. This was distributed before the second batch of emails was sent.

Discussion among the pSCANNER team highlighted the importance of centralized coordination in accomplishing this study efficiently and effectively. First, one site tracked coordinated the project, created template invitations and letters, tracked responses and communicated with all site project managers and with the telephone survey firm. There was one contract for the survey firm that covered all sites. A different site developed and hosted the Qualtrics survey, delivering unique survey links that each site could match to their participant records. This allowed for efficient site-level tracking while allowing patient privacy to be maintained at the local site.

Patients’ willingness to participate in clinical studies is critical for advancing precision medicine research. As we learn more about patients’ attitudes and preferences toward enrolling in research, the information can be useful in designing future studies and improving patient recruitment and retention. Overall, the majority of patients in our survey showed favorable attitudes towards participating in future research studies. High interest level was shown for one-time completion of survey/questionnaire or one-time submission of a biospecimen. On the other hand, it is not surprising that respondents were less enthusiastic about participating in research that involved more invasive, higher burden processes, or research that involved family members. This is in line with other studies that investigated factors influencing clinical research participation.\textsuperscript{16,17} The high level of hypothetical participation among respondents maybe due to the fact that the respondents were already enrolled in this survey study. Also, respondents might already have established trust in research conducted by state-supported, large research universities and their academic medical centers.

While we did find a significant association between participant race and interest in all research activities, we were not able to make interpretations on each race factor and the research interests. Some studies have reported that racial/ethnic minorities tend to have reduced willingness to participate in research due to skepticism or fear toward medical research.\textsuperscript{8,18,19} On the contrary, other studies have shown that participants with minority identity showed greater willingness to participate in medical research.\textsuperscript{15,20} Prior studies have also reported that women, elderly, and minorities were consistently under-represented in cardiovascular disease (CVD) clinical trials.\textsuperscript{21-23} However, more recent online survey of 504 respondents with CVD showed that there were no significant differences in willingness to participate in clinical research based on sex.\textsuperscript{24} Further analysis documented that willingness to participate in clinical research were similar among patients with CVD and patients with other chronic conditions.\textsuperscript{24} Unlike our study which assessed participants’ willingness to participate in different types of research activities, this survey used a single-item to measure willingness to participate in clinical research. Cohort studies must be able to recruit participants who are representative of the population in order to be meaningful.

The majority of our survey respondents preferred to be contacted by research staff or their physician by personal phone call for future trial participation. This finding is similar to other studies reporting a great willingness to participate in research if asked by their own doctor.\textsuperscript{15} Although still at an early stage, using social media as a tool for clinical trial recruitment has shown some promising results in several clinical trials;\textsuperscript{25,26} even showing effectiveness in reaching historically hard-to-reach populations.\textsuperscript{27,28} Interestingly, social media was the least preferred means of contact among our survey respondents. This may be because our mean age of respondents were 63 years old and they may not embrace social media platforms as much as their younger counterparts. Future studies that assess age and cultural differences in the use of social media for recruitment tool may provide additional insights.

Several limitations must be noted. First, the survey respondents were limited to heart failure patients who were primarily treated in academic healthcare facilities in California. Thus, given geographic and cultural differences, these findings may not be generalizable to patient populations in other health settings or regions of the United States. Additionally, this was a self-report survey and therefore data regarding willingness to participate in research may be prone to social desirability biases. It is possible that participants reported higher willingness for future research participation to make a favorable impression. Furthermore, the cross-sectional design of this study makes it difficult to draw casual inferences about factors affecting respondents’ willingness to participate in research. Despite these limitations, our results suggest that CDRN infrastructure is able to support prospective enrollment of patients into future clinical studies.
Conclusion

The willing participation of patients is a critical element in ongoing national efforts to construct cohorts for precision medicine and other health research. However, enrolling and retaining patients to clinical research studies can often be challenging. This survey demonstrates that CDNs can be a source for recruiting for large study cohorts among a geographically diverse population and engaging patients into participating in future studies. Yet, identification with the local institution and investigators may be an important element in garnering patient willingness to engage in future research. Overall, most of the survey respondents showed interest in participating in future studies. The lessons learned in this CDN may be useful for other cohort studies and network-based research efforts.

Acknowledgments

This work was partially supported by a Patient-Centered Outcomes Research Institute (PCORI) Award (Contract CDN-1306-04819, LO-M, KK, DM). The statements in this article are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee. This study was also supported by the UCD Clinical and Translational Science Center grant UL1TR001860 and the KL2 scholar program supported by the National Center for Advancing Translational Sciences; NIH UL1 TR000002 (to JEL).

Other members of the pSCANNER team involved in this study included: Hugo O. Campos, Patient Co-Investigator; UC Davis: Michael Hogarth, Kate Marie, Calvin Chang, Sarah C. Haynes; UC San Diego: Michele Day, Paulina Paul, Rita German-Kurtz; UC Irvine: Pietro Galassetti, Lisa Dahm, Robynn Zender, Roni Bracha, Ayan Patel; UC Los Angeles: Doug Bell, Marianne Zachariah; UC San Francisco: Mary Whooley, Rhiannon Croci; VA San Diego: Dena Rifkin, Carl Stepnowksy, Carolina Smales; USC: Jason Doctor, Tara Knight.

References


Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning

Olivia Choudhury\textsuperscript{1}, PhD, Yoonyoung Park\textsuperscript{1}, ScD, Theodoros Salonidis\textsuperscript{2}, PhD, Aris Gkoulalas-Divanis\textsuperscript{3}, PhD, Issa Sylla\textsuperscript{1}, BA, Amar K. Das\textsuperscript{1}, MD, PhD

\textsuperscript{1}IBM Research Cambridge, Massachusetts, USA
\textsuperscript{2}IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
\textsuperscript{3}IBM Watson Health, Cambridge, Massachusetts, USA

Abstract

Using electronic health data to predict adverse drug reaction (ADR) incurs practical challenges, such as lack of adequate data from any single site for rare ADR detection, resource constraints on integrating data from multiple sources, and privacy concerns with creating a centralized database from person-specific, sensitive data. We introduce a federated learning framework that can learn a global ADR prediction model from distributed health data held locally at different sites. We propose two novel methods of local model aggregation to improve the predictive capability of the global model. Through comprehensive experimental evaluation using real-world health data from 1 million patients, we demonstrate the effectiveness of our proposed approach in achieving comparable performance to centralized learning and outperforming localized learning models for two types of ADRs. We also demonstrate that, for varying data distributions, our aggregation methods outperform state-of-the-art techniques, in terms of precision, recall, and accuracy.

1 Introduction

Adverse Drug Reactions (ADRs) are a major concern for medical practitioners, healthcare system, and pharmaceutical industry. As patients can experience expected and sometimes unexpected negative outcomes from taking any drug, delayed detection of ADRs can pose life-threatening risks to patients; posing considerable legal, financial, and social repercussions to the manufacturing companies and regulatory agencies. The use of medical data, such as claims and electronic health records (EHR), has become common in providing rich insights on health services and supporting ADR investigation\textsuperscript{1}. Advancements in machine learning and artificial intelligence have produced a number of analytic methods that can be applied to such high-dimensional data for the purpose of predicting adverse reactions\textsuperscript{2}. However, making timely and accurate predictions remains a challenge. Due to the distributed nature of healthcare data, obtaining a sufficiently large dataset to detect rare events requires merging data from different data silos. Analyses generated from different data sources can be conflicting or imprecise\textsuperscript{3}, necessitating methods to appropriately aggregate results.

Prior work to resolve these issues often have limitations in their approach. The Food and Drug Administration (FDA)'s self-report Adverse Event Reporting System (FAERS)\textsuperscript{4}, collects ADR data into a traditional, centralized database. A single database approach is the most straightforward way to explore ADRs, but information owned by different entities is seldom shared due to significant privacy concerns. Moreover, creating and maintaining such a large data repository incurs resource and system-level constraints, including high latency and single points of vulnerability (failure, breach). To avoid such overhead and risks, the FDA created the Sentinel system to monitor the safety of its regulated products, using a distributed data network\textsuperscript{5,6}. The network comprises multiple stakeholders, each maintaining a large claims database. Despite the distributed framework and large-scale data amassed from active participation of data partners, Sentinel has limited analytic capabilities. Limitations of other state-of-the-art systems include access to potentially small-scale, sparse, and low-quality hospital records\textsuperscript{7}. In addition, current claims-based frameworks experience a time lag between ADR instance, claim submission, adjudication, and consolidation of the claim into a database. EHR data, collected in near real-time, is therefore a promising alternative, but comes with the aforementioned quality concerns. Hence, there is an unmet need for accurate, scalable, and efficient solutions for predicting ADRs using distributed health data, that also protects the privacy of patients.

To address this challenge, we present a federated learning-based framework that permits health data to be distributed across multiple sites. Federated learning\textsuperscript{8} has brought a paradigm shift in the construction of machine learning models from distributed data sources maintained by various organizations. Under such a decentralized, collaborative learning
setting, each site contributes to the computation of a global model while simultaneously shielding its own data from leakage to distrusted third parties. Our framework allows us to train a global model based on each site’s local data, without ever moving the raw data from their respective sites. To the best of our knowledge, this is the first implementation of federated machine learning algorithms that leverages distributed electronic health data for predicting ADRs. ADR prediction itself brings significant challenges for federated learning due to the huge imbalance between the majority class of individuals who do not suffer from ADRs, and the minority class of individuals with severe ADRs. To address this issue, we propose two novel methods of aggregating model updates from the sites and compare their performance with that of the state-of-the-art alternative. To show the effectiveness of our proposed approach, we consider two use cases: (i) prediction of chronic opioid usage for patients taking opioid drugs, and (ii) prediction of extrapyramidal symptoms for patients taking antipsychotic drugs. We conduct a comprehensive experimental evaluation using real-world patient data.

The key contributions of our work include: (1) implementing federated models for ADR prediction based on three supervised learning algorithms; (2) proposing and implementing two novel methods of aggregating local model updates in a federated setup; (3) demonstrating the effectiveness of our approach in analyzing sensitive, distributed, and highly imbalanced real-world electronic health data; (4) conducting a comparative analysis to evaluate our approach against state-of-the-art alternatives; and (5) demonstrating scalability of our approach for varying number of sites, data size, and data distribution.

2 Background

The sensitive and distributed nature of electronic health information in real-world scenarios motivate the need for a mechanism that can learn from data residing in silos, while accounting for data privacy. This compels us to explore the potential and value of federated learning for ADR prediction. Federated learning enables training a global model from distributed data, without having the sites exchanging any raw sensitive data. The global model is distributed to each site, where an instance is trained locally. The updates from locally trained instances are then aggregated to improve the global model, which is shared again with the sites for another round of training. This iterative process, illustrated in Figure 1, terminates when a performance criterion is met.

Initial implementations of federated learning were intended for image classification and language modeling on mobile devices. Existing literature aims to improve the performance of deep networks in a federated setting. There is currently very limited research focusing on the application of federated learning in healthcare. Recent work noted the effectiveness of federated models in predicting hospital admissions using EHR data. However, the potential of federated learning in healthcare applications that make use of claims or EHR data for ADR prediction is yet to be explored. Moreover, the existing method of aggregating updates from local models relies on the size, rather than the inherent characteristics, of the data. This approach may not work well in healthcare applications, which often deal with skewed, sparse, and imbalanced datasets. Hence, exploring the underlying characteristics of distributed data to improve the predictive capability of the global model is also an important research direction.

Unlike the methods that focus on surveillance of all potential ADRs for a given drug, specific prediction typically employs supervised learning algorithms. Commonly used algorithms are logistic regression, random forest, decision trees, Support Vector Machine (SVM), and neural networks. Prior works on ADR prediction with machine learning methods are largely limited to centralized models, where all data are available to the researcher in a centralized data store. A majority of these works also lack evaluation on real-world datasets. For instance, distributed logistic
regression based on multi-party computation, was studied using simulated data.18

In this paper, we implement federated models based on three supervised classification algorithms: SVM, single-layer perceptron, and logistic regression, using stochastic gradient descent (SGD)-based optimization, which provides a generic approach for the algorithms to learn local models and aggregate their parameters to improve the global model and is the method currently supported by federated learning.

3 Methods

3.1 Data and cohort selection

To evaluate our approach, we used the Limited IBM MarketScan Explorys Claims-EMR Data Set (LCED). The data is procured from administrative claims and EHR data of over 1 million commercially insured patients over 5 years (2012-2017), each varying in lengths of medical activity. LCED contains patient-level features, such as demographics, diagnostic codes, outpatient prescription fills, laboratory results, and inpatient admission records. From this dataset, we defined our cohort based on the occurrences of two drug-ADR pairs during a 5 year period. Our selection was based on multiple factors: the number of patients taking the drugs of interest, the prevalence of the specific ADR of interest, and the feasibility of reliably detecting the ADR incidents using the LCED data.

Use case I: The first use case regards predicting chronic opioid usage (potential opioid use disorder) among patients who receive an opioid drug. Opioid abuse is currently one of the most pressing public health issues in the United States. The first exposure to opioids, for many who develop an abuse disorder, is a physician’s prescription to alleviate symptoms of another condition. We closely follow Zhang et al19 to construct the opioid user cohort by first identifying all patients who had one or more opioid prescriptions. We define opioid discontinuation as having 180 or more opioid-free days during one year (365 days) after the initial prescription, and chronic users of opioid as those who did not discontinue during the same period. Opioid-free day is calculated by subtracting from 365 days the sum of all subsequent opioid prescriptions’ days-supply during the year following initiation. We excluded patients who had a diagnosis of opioid use disorder, cancer, or received hospice service in the year prior to the initiation, as their chronic opioid usage was expected. After identifying a number of risk factors for chronic opioid use from the literature, we hand-generate them from the data as predictive features. Examples include age, gender, smoking and alcohol habits, diagnoses of different types of pain (spinal injury, arthritis, etc.), surgical procedures, use of psychotropic drugs, and the morphine milligram equivalent (MME) of the initial opioid prescription normalized by its days-supply.

Use case II: The second use case regards predicting incidents of extra-pyramidal symptoms (EPS), a type of motor dysfunction, from the use of antipsychotic drugs. Broad off-label use of antipsychotics is well known despite the risk for developing ADRs like EPS. Similar to the opioid cohort, we start from identifying patients who had one or more prescription for an antipsychotic. We define the incidence of EPS following antipsychotic initiation using International Classification of Diseases (ICD) Ninth and Tenth Revision diagnosis codes. Examples of patient level predictive features identified from the literature include age, gender, smoking and alcohol habits, type of the first antipsychotic (first vs. second generation), other drug use (other psychotropic drugs, drugs related to Parkinson’s Disease, etc.), comorbid diagnoses (dementia, psychiatric diagnoses, epilepsy, Parkinson’s Disease, etc.), and number of prior hospital admissions. We generate these features using the data from the one-year period prior to the initiation of antipsychotic for each patient.

<table>
<thead>
<tr>
<th>Cohort</th>
<th># Patients</th>
<th>Mean Age (SD)</th>
<th>Female (%)</th>
<th># Patients with ADR</th>
<th>Class ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opioid</td>
<td>1,161,048</td>
<td>48.6 (18.2)</td>
<td>57.8</td>
<td>69,863</td>
<td>1:15.6</td>
</tr>
<tr>
<td>Antipsychotic</td>
<td>86,674</td>
<td>50.1 (21.7)</td>
<td>60.5</td>
<td>1,314</td>
<td>1:65</td>
</tr>
</tbody>
</table>

Table 1: Summary of the two drug-ADR pair cohorts. SD: Standard deviation

Table 1 provides summary statistics of the two cohorts. The imbalance between the majority (non-ADR) and minority (ADR) classes was consistent with our expectation, since severe ADRs are rare with most of the drugs available on the market. Our primary analyses focus on the opioid cohort, where a better balance is achieved with a larger sample size. We examine the antipsychotic cohort as an example of small, highly imbalanced data that one can encounter frequently in the healthcare domain. The implication of severe class imbalance is discussed in the following sections.
3.2 ADR prediction model

Consider a general binary classification problem, where features, denoted by $x_k$ (for the $k^{th}$ feature), are drawn from a feature space $X$. The corresponding labels $y_k$ are drawn from the label space $Y := \{-1, 1\}$. Let the features corresponding to positive labels be denoted by $X_+$ and those corresponding to negative labels by $X_-$, that is

\[ X_+ = \{ x_k \in X : y_k = +1 \} \quad \text{and} \quad X_- = \{ x_k \in X : y_k = -1 \} \]

For any $x_k^+ \in X_+$ and $x_k^- \in X_-$, the objective of binary classification is to construct a function $f : X \rightarrow Y$ such that

\[ f(x_k^+) = +1 \quad \text{and} \quad f(x_k^-) = -1. \]

In this paper, we denote cases of ADR as labels $y_k = +1$, and cases of non-ADR as $y_k = -1$.

3.2.1 Cost-sensitive learning

Class imbalance is intrinsic to ADR prediction. Since most classification algorithms assume balanced class distributions or equal misclassification costs, they fail to represent the characteristics of imbalanced data and are more likely to classify new observations to the majority class. For ADR prediction, the cost of a false negative classification should be much higher than that of a false positive classification. Recent work on imbalanced learning can be categorized into sampling methods, cost-sensitive methods, and active learning methods. As discussed in sampling methods, such as undersampling the majority class or oversampling the minority class, either discard potentially useful data or can lead to overfitting. Since our dataset does not comprise unlabeled samples, active learning is not applicable.

Hence, to mitigate the challenge of skewed data distribution, we incorporate cost-sensitive learning, wherein we increase the cost associated with misclassifying a minority class sample. Specifically, if $C_{FN}$ and $C_{FP}$ denote the cost of a false negative and false positive in a cost matrix, respectively, then we set $C_{FN} > C_{FP}$. The magnitude of cost depends on the problem at hand and we determine their values using grid search.

3.2.2 Centralized model

For the purpose of binary classification of samples into ADR and non-ADR cases, we consider three supervised classification methods: SVM, single-layer perceptron, and logistic regression. We implemented these algorithms using scikit-learn version 0.20.2. To establish benchmark results, we first evaluate the performance of the classifiers in a centralized learning approach. This represents the scenario of gathering data from multiple sites for training a machine learning model. For each cohort, we split its entire dataset into two parts: 70% for training, and 30% for testing, where $X_{\text{train}}$ and $Y_{\text{train}}$ denote the feature and label sets for training, and $X_{\text{test}}$ and $Y_{\text{test}}$ denote the feature and label sets for testing. As the splits are stratified, the proportion of positive and negative cases in each split is the same as the entire dataset. After standardizing the features, we use 5-fold cross-validation to train the models on $X_{\text{train}}$ and $Y_{\text{train}}$, and test them on $X_{\text{test}}$. To incorporate cost-sensitive learning, we update the class_weight parameter in scikit-learn based on class frequencies.

3.2.3 Localized model

Since healthcare and biomedical data is rife with sensitive information, sharing such data across sites or transferring it to a centralized database is often restricted. In such cases, a site has to rely on its own data for predictive analytics. We consider this scenario while designing localized models for ADR prediction. We train each classifier on a site’s data, without leveraging data from other sites. Let us suppose there are $N$ sites, representing hospitals or data owners. We use horizontal partitioning to split the training data into $N$ disjoint subsets. We partition $X_{\text{train}}$ into $\{X_{\text{train}}^i\}_{i=1}^N$, where $\bigcup_{i=1}^N X_{\text{train}}^i = X_{\text{train}}$ and $X_{\text{train}}^i \cap X_{\text{train}}^j = \emptyset$, $\forall i, j \in \{1, ..., N\}$, for $i \neq j$. We follow the same logic to partition the corresponding label set $Y_{\text{train}}$ into $\{Y_{\text{train}}^i\}_{i=1}^N$. In the case of localized learning, the classifiers are trained on a single site’s data $\{X_{\text{train}}^i\}_{i=1}^N$ and $\{Y_{\text{train}}^i\}_{i=1}^N$, and tested on $X_{\text{test}}$. The limited availability of data may fail to account for detection of rare events. We consider the results obtained from the localized models for benchmark analysis with federated and centralized learning models.
3.2.4 Federated model

In this paper, we focus on classification models that can be trained using gradient descent optimization, as currently supported by federated learning. Similarly to the scenario of localized model, for $N$ sites, we randomly partition the training data into $N$ disjoint subsets of feature set $\{X_{i,\text{train}}\}_{i=1}^N$ and corresponding label set $\{Y_{i,\text{train}}\}_{i=1}^N$. Let $T$ denote the rounds of aggregating local model updates. For stochastic gradient descent, let $\eta$, $E$, and $Batch$ denote the learning rate, number of epochs, and batch based on a given batch size $B$, respectively. Let $F_i(w)$ be the local loss function of the $i^{th}$ site with respect to its model parameter $w$. As described in Section 2, a global model is shared with each site, which trains the model on its local data. During local model training, based on given $\eta$, $E$, and $Batch$, at each site, we compute average gradient $(\nabla F_i(w))$ with respect to its current model parameter $w$. We then compute weighted average to aggregate the parameter updates from the local models. The process is repeated until a convergence criterion, such as minimization of loss function, is satisfied. The process of training the global model only relies on updates from the local models, rather than raw data residing at the sites. Algorithm 1 presents the core algorithm of federated learning, where the weight $w_D^i$, used to compute weighted average, depends on the model aggregation method used. We first implement and evaluate the performance of the state-of-the-art model averaging approach, known as federated averaging. It computes a weighted average based on the fraction of data residing at each site. For this case, the weight $w_D^i$ is equal to $\frac{|D_i|}{|D|}$, where $|D_i|$ and $|D|$ denote the size of data at the $i^{th}$ site and the entire dataset, respectively. Such an approach may fail to consider the inherent characteristics of data distribution at the sites. For the use case of ADR prediction, federated averaging would not account for imbalanced data and the varying distribution of ADR cases across sites. Since such scenarios are common when dealing with real-world health data, particularly in predicting rare events, it is important to explore other aggregation approaches.

3.2.5 Aggregation of local model updates

In this paper, we propose two novel methods of aggregating local model updates. The first method is particularly designed for training data with imbalanced classes. For each site, we estimate the class ratio of its training data to assign a corresponding weight, as denoted by $w_D^i$. This would imply that sites with cases of rare events, would have higher impact when improving the global model. For the second approach, we consider loss per sample, the change in the loss function during local model training. Since a gradient descent-based method attempts to minimize the loss function, we determine its rate of convergence. This is measured by the metric epoch, which is the maximum number of passes over the training data until convergence. Based on each site’s epoch and training data size, we assign a weight, corresponding to $w_{Ep}^i$, for future aggregation. Using this approach, sites which require less training samples to reach convergence faster, will be assigned a higher weight during aggregation.

To evaluate these methods, we create a separate partition of the training data, based on the opioid cohort, to represent unequal distribution of class labels, as shown in Table 2. We do not conduct the same experiment with the antipsychotic cohort due to the limited number of minority class labels (ADR).

<table>
<thead>
<tr>
<th>Site #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td># ADR</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td># Non-ADR</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>100,000</td>
<td>150,000</td>
<td>200,000</td>
<td>250,000</td>
<td>16,258</td>
<td>16,258</td>
<td>16,258</td>
</tr>
<tr>
<td>Class Ratio</td>
<td>1:1</td>
<td>1:1</td>
<td>1:1</td>
<td>1:20</td>
<td>1:30</td>
<td>1:40</td>
<td>1:50</td>
<td>1:3.2</td>
<td>1:3.2</td>
<td>1:4.1</td>
</tr>
</tbody>
</table>

Table 2: Partitioning of the opioid cohort training data with varying class ratio.

4 Experimental Evaluation

In this section, we present experimental results to evaluate our proposed approach. We discuss the evaluation metrics we used in this study, followed by a comparative analysis to demonstrate the effectiveness of the proposed system.

4.1 Evaluation metrics

To measure the predictive capability of the centralized, localized, and federated learning models, we compute precision, recall, and accuracy scores. As noted in prior work\textsuperscript{21,26}, precision and recall are better indicators for models
dealing with imbalanced data. We also report the runtime incurred in training the models for each setup. All experiments were run on an Intel(R) Xeon(R) E5-2683 v4 2.10 GHz CPU equipped with 16 cores and 64 GB of RAM.

Algorithm 1 Federated Learning Model for ADR Prediction

1: function \textsc{UPDATEGLOBALMODEL}
2: \hspace{1cm} initialize $w_0$
3: \hspace{1cm} for $t = 1$ to $T$ do
4: \hspace{2cm} for $i = 1$ to $N$ do
5: \hspace{3cm} $w^i_{t+1} = \textsc{UPDATELOCALMODEL}(i, w_t)$
6: \hspace{2cm} $w_{t+1} = \sum_{i=1}^{N} w^i_D * w^i_{t+1}$

7: function \textsc{UPDATELOCALMODEL}(i, w)
8: \hspace{1cm} for $e = 1$ to $E$ do
9: \hspace{2cm} for $b \in \text{Batch}$ do
10: \hspace{3cm} $w = w - \eta \nabla F_i^b(w)$
11: \hspace{1cm} return $w$

For the set of experiments comparing centralized, localized, and federated learning models, we examine the observed differences in performance metrics in two ways: (a) by calculating the % relative error of federated learning and localized learning with respect to centralized learning when using federated averaging\(^9\), and (b) by testing the statistical significance of the difference using the Wilcoxon signed-rank test at 0.05 significance level.

4.2 Comparative analysis

We compare precision, recall, and accuracy of the federated learning (FL) model with two benchmark models: centralized learning (CL) and localized learning (LL). We execute the experiments 10 times for each setup. Figures 2, 3, and 4 report these metrics for the two datasets. As seen in Figure 2, SVM and perceptron yield similar scores and perform better than logistic regression. For all classifiers and datasets, federated learning exhibits comparable performance with respect to centralized learning. At the same time, due to the lack of sufficient training data, localized models do not perform well. It must be noted that the precision score for the antipsychotic data is higher than that for the opioid data. This is due to having a lower number of false positives, possibly because the former dataset is severely imbalanced with a class ratio of 1 : 65.

Figure 3 presents the recall scores of the models for the two datasets. Perceptron generated the highest recall score, followed by SVM and logistic regression. Federated learning performed as good as centralized learning, and outperformed the localized learning models. Since the implementation of cost-sensitive learning reduced the cases of false negative, even with such imbalanced data, centralized and federated learning models for SVM and perceptron achieved high recall.

Figure 4 compares the accuracy score of the three models for the given dataset. Similarly to previous observations, Perceptron and SVM perform better than logistic regression. Our federated learning approach achieves comparable and better accuracy than centralized and localized learning models, respectively.

We observe that the % relative error values from federated learning are smaller than those from localized learning (Table 3). To put the numbers into a context, a difference of 5% in recall can translate to missing 5 out of 100 ADR cases compared to using centralized learning. Higher recall is desirable given the potential cost of missing severe ADR cases, and therefore federated learning with low % relative error is preferred. Based on statistical testing, in both opioid and antipsychotic data, the performance of centralized learning and federated learning is comparable for all three metrics for all classifiers. On the other hand, the performance of localized learning is inferior compared to either centralized or federated learning for the three evaluated metrics (all p values < 0.05).

In Table 4, we report the running time (in seconds) incurred in training the models for different setups. As expected, centralized learning requires a lot of time as it involves training the models on the entire training dataset. Federated learning requires significantly less time to train the models. Localized learning models train on a subset of the data on a single round, due to which they incur the lowest running time. For both datasets, perceptron required higher running time, compared to SVM and logistic regression. Due to the considerably large scale of opioid data, it consistently required more time to train the models.

To demonstrate the scalability of federated learning models, we further measure their predictive capability, in terms of precision and recall, for a varying number of sites and data sizes. As the number of sites increases, the size of training data residing at each site proportionally decreases. Due to the imbalanced nature of the data, this has a pronounced
impact on the recall score, as evident in Figure 5. This scenario also accounts for the ability of the system to handle varying sizes of training data.

As previously discussed, we partition the opioid cohort such that the sites have a varying distribution of ADR and non-ADR cases (see class ratios in Table 2). We compare the effectiveness of our two proposed aggregation methods, in terms of precision, recall, and accuracy, with respect to default averaging (without weights) and federated averaging (based on data size). As seen in Table 5, for all evaluation metrics, our methods, particularly aggregation based on loss per sample, outperforms the state-of-the-art method of aggregation. This result implies that for skewed datasets, it is very important to consider the underlying characteristics of the data when aggregating local models.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>FL vs CL</th>
<th>LL vs CL</th>
<th>FL vs CL</th>
<th>LL vs CL</th>
<th>FL vs CL</th>
<th>LL vs CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opioid</td>
<td>SVM</td>
<td>2.84(1.64)</td>
<td>6.44(1.59)</td>
<td>3.10(2.58)</td>
<td>8.49(3.88)</td>
<td>3.96(2.33)</td>
<td>13.34(2.50)</td>
</tr>
<tr>
<td></td>
<td>Perceptron</td>
<td>1.11(.73)</td>
<td>7.16(2.49)</td>
<td>7.32(5.45)</td>
<td>9.06(6.06)</td>
<td>5.31(3.99)</td>
<td>12.88(4.58)</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>1.86(1.11)</td>
<td>11.28(2.56)</td>
<td>3.21(2.81)</td>
<td>11.55(3.37)</td>
<td>2.66(2.54)</td>
<td>12.55(4.49)</td>
</tr>
<tr>
<td>Antipsychotic</td>
<td>SVM</td>
<td>1.61(1.04)</td>
<td>11.82(2.28)</td>
<td>2.24(1.49)</td>
<td>15.71(2.55)</td>
<td>4.47(3.63)</td>
<td>16.45(2.50)</td>
</tr>
<tr>
<td></td>
<td>Perceptron</td>
<td>2.81(1.68)</td>
<td>12.73(3.53)</td>
<td>2.39(1.95)</td>
<td>14.08(5.00)</td>
<td>4.87(2.50)</td>
<td>11.22(4.19)</td>
</tr>
<tr>
<td></td>
<td>LogReg</td>
<td>1.42(1.26)</td>
<td>11.13(2.05)</td>
<td>2.21(1.99)</td>
<td>12.70(4.60)</td>
<td>4.19(3.90)</td>
<td>7.90(6.13)</td>
</tr>
</tbody>
</table>

Table 3: Comparison of relative error (%) for federated learning (FL) and localized learning (LL) with respect to centralized learning (CL). The values denote average (standard deviation) over 10 iterations.

<table>
<thead>
<tr>
<th></th>
<th>Opioid</th>
<th>Antipsychotic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CL</td>
<td>FL</td>
</tr>
<tr>
<td>SVM</td>
<td>612.8 (8.5)</td>
<td>122.2(3.4)</td>
</tr>
<tr>
<td>Perceptron</td>
<td>842.8 (9.0)</td>
<td>117.6(2.9)</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>513.7 (6.4)</td>
<td>102.7(3.4)</td>
</tr>
</tbody>
</table>

Table 4: Time (in seconds) incurred in training the centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression. The times denote average (standard deviation) over 10 iterations.

Figure 2: Comparison of precision score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with (a) opioid data and (b) antipsychotic data.

5 Discussion

The availability of electronic health data brings countless opportunities to investigate and predict ADR, provided that the hurdles in gathering and using such data are overcome. In this work, we proposed and evaluated the use of federated
Figure 3: Comparison of recall score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with (a) opioid data and (b) antipsychotic data.

Figure 4: Comparison of accuracy score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with (a) opioid data and (b) antipsychotic data.

Learning to address the limitations of ADR prediction frameworks based on centralized learning. We demonstrated that SVM and perceptron perform better than logistic regression with respect to precision, recall, and accuracy. Perceptron has higher recall values, making it the preferable classifier for ADR detection, where false negatives generally have more significant consequences than false positives. We also demonstrated that the performance of federated learning models is comparable to that of centralized learning, implying that a federated learning framework can be used to predict ADR without compromising the model performance, while bypassing the challenges associated with centralized learning. An important finding of our evaluation regards the quality of our proposed aggregation approach with loss to sample ratio weighting, which achieves superior performance compared to state-of-the-art federated averaging. This approach is advantageous in federated learning applications with real-world health data, where severe class imbalance is a norm, rather than an exception.

In this paper, we focused on classification algorithms that are amenable to distributed solution using gradient descent, as currently supported by the federated learning paradigm. In the future, we plan to extend our federated learning framework to other types of algorithms, such as decision trees and gradient boosting, as well as applications where large-scale distributed datasets are common and deep learning models are applicable. We will leverage other charac-
Figure 5: Effect of varying number of sites on precision and recall scores of federated learning models (SVM, perceptron, logistic regression) with (a) opioid data and (b) antipsychotic data.

<table>
<thead>
<tr>
<th>Aggregate</th>
<th>SVM Precision</th>
<th>Perc Precision</th>
<th>LR Precision</th>
<th>SVM Recall</th>
<th>Perc Recall</th>
<th>LR Recall</th>
<th>SVM Accuracy</th>
<th>Perc Accuracy</th>
<th>LR Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>.93 (.01)</td>
<td>.91 (.02)</td>
<td>.91 (.01)</td>
<td>.63 (.02)</td>
<td>.68 (.02)</td>
<td>.59 (.01)</td>
<td>.64 (.02)</td>
<td>.68 (.03)</td>
<td>.59 (.01)</td>
</tr>
<tr>
<td>Fed Avg</td>
<td>.93 (.02)</td>
<td>.91 (.01)</td>
<td>.90 (.02)</td>
<td>.58 (.03)</td>
<td>.64 (.01)</td>
<td>.54 (.02)</td>
<td>.58 (.01)</td>
<td>.64 (.02)</td>
<td>.54 (.02)</td>
</tr>
<tr>
<td>Class ratio</td>
<td>.94 (.01)</td>
<td>.92 (.01)</td>
<td>.90 (.01)</td>
<td>.72 (.02)</td>
<td>.68 (.01)</td>
<td>.61 (.02)</td>
<td>.71 (.01)</td>
<td>.67 (.01)</td>
<td>.62 (.01)</td>
</tr>
<tr>
<td>Loss/sample</td>
<td>.94 (.01)</td>
<td>.92 (.01)</td>
<td>.91 (.01)</td>
<td>.75 (.02)</td>
<td>.69 (.01)</td>
<td>.63 (.01)</td>
<td>.74 (.01)</td>
<td>.69 (.01)</td>
<td>.63 (.02)</td>
</tr>
</tbody>
</table>

Table 5: Comparison of our proposed aggregation methods (based on class ratio and loss/sample) with respect to averaging and federated averaging methods. For SVM, perceptron (Perc), and logistic regression (LR), we report the average (standard deviation) values of precision, recall, and accuracy scores.

We will also explore potential approaches for tuning hyperparameters of the global model in a federated setup. We intend to work on approaches for privacy-preserving federated learning, which protect patients’ privacy against adversarial attacks, in addition to not exchanging raw data while training the models.

References
6. Qoua L. Her, Jessica M. Malenfant, Sarah Malek, Yury Vilk, Jessica Young, Lingling Li, Jeffery Brown, and Sengwee Toh. A Query Workflow Design to Perform Automatable Distributed Regression Analysis in Large Distributed Data Networks. eGEMs, 2018.


Leveraging Clinical Expertise as a Feature - not an Outcome - of Predictive Models: Evaluation of an Early Warning System Use Case

Sarah Collins Rossetti, RN, PhD, FACMI, FAMIA1,2, Chris Knaplund, MPhil1, Dave Albers, PhD1, Abdul Tariq PhD2, Kui Tang3, David Vawdrey PhD1,3, Natalie H. Yip, MD4, Patricia C. Dykes, RN, PhD5,6, Jeffrey G. Klann PhD6, Min Jeoung Kang RN, PhD5,6, Jose Garcia5, Li-Heng Fu, MD1, Kumiko Schnock, RN, PhD5,6, Kenrick Cato, RN, PhD2,3

1Columbia University, Department of Biomedical Informatics, New York, NY; 2Columbia University, School of Nursing, New York, NY; 3New York Presbyterian Hospital, New York, NY; 4Columbia University Medical Center, New York, NY, 5Brigham and Women’s Hospital, Boston, MA; 6Harvard Medical School, Boston, MA

Abstract

Identifying patients at risk of deterioration in the hospital and intervening more quickly to prevent adverse events is a top patient safety priority. Early warning scores (EWS) identify at risk patients, but there is much opportunity for improvement particularly related to increasing lead time – the time from an alert trigger to adverse event (e.g., cardiac arrest, death). Our team develops healthcare process models of clinical concern (HPM-CC) and in this work has identified documentation signals that are proxies of nurses concern and can be used to predict patient risk earlier than current EWS systems that rely only on physiological data. We compared the performance of a validated EWS - the MEWS - to our novel model (MEWS-CC) comprised of MEWS criteria plus 3 proxy variables of nursing concern. MEWS-CC performed similarly to MEWS, with the added benefit of increased time from EWS trigger to event by 5-26 hours.

Introduction

Identifying preventable deterioration in the hospital setting is a patient safety priority with suboptimal outreach interventions and outcomes to date.1 While the exact estimates of preventable deaths and adverse events in the hospital setting have been debated, it is widely accepted that the numbers are much too high.2-4 Predicting a patient’s future state is a great challenge and focus of early warning score (EWS) models to identify patients at risk of deterioration in the hospital setting. EWS models predominately use vital sign abnormalities to predict cardiac arrest and death within a set time period, such as 24 or 48 hours, and overall have good predictive performance but there is insufficient evidence of positive impact on patient outcomes once implemented in the clinical setting.5,5 Many EWS studies report on the ability to predict the event at any point using the prior 24 or 48 hours, rather than the more useful clinical objective of maximizing the time from first alert trigger to event in order to enable sufficient and effective clinical intervention.5,6 Further, most studies ignore the clinically established criteria of “concern” or “worried about the patient”7 A 2014 review of EWS found that only 3 of 13 unique models included a concept of clinical concern where the clinician, typically the nurse, answered structured questions if he/she were subjectively concerned of the patient.5,7,8 In this era of big data and artificial intelligence, it is noteworthy to recall that prediction or projection of a patient state occurs routinely by clinicians in clinical practice and the predominant reason to activate a rapid response team (RRT) was the subjective criteria of “concern” before EWS criteria changes in vital sign parameters were reached.9-11 Machine learning is limited by availability of computable data. However, clinicians consume vast amounts of information during the course of clinical care that are never recorded in the electronic health record (EHR), such as by looking at or talking to a patient. The breadth of data that clinicians are exposed to and observe become information that triggers clinicians to record data in the EHR that they deemed clinically relevant and salient to the patient’s care. Therefore, the presence of recorded data itself is relevant and available to machine learning algorithms.

The most common physiological criteria used to trigger a RRT are low oxygenation, low systolic blood pressure, change in level of consciousness, or change in heart rate or respiratory rate (RR).12-14 Respiratory rate has also been shown both as the most predictive vital sign for adverse events and a variable with significant interactions with other vital signs in EWS models.15 However, delays in activation of RRTs have been attributed to overreliance on waiting for physiological criteria to be met, failure to identify deterioration, and complexities in workflow and team communication.16 Clinically intervening on delays in activation of RRTs is a priority. In this study we demonstrate a
method to measure “concern” or “worry about a patient” for use in EWS that can in turn be validated and implemented in the clinical setting.

Clinical expert intuition – or prediction of patient trajectories - is generally understood as complex decision making in which data and information are filtered and synthesized for expert pattern matching which leads to skilled judgements based on possible unconscious observations or inability to articulate the cues that guide them.\textsuperscript{16,17} Aligned with this notion of clinician decision making informed by trends to predict trajectories, one EWS study created variables that track the trend of changes in each vital signs (24hrs mean, variance, max, min) and showed that the model performed better in predicting outcomes then the models that only take into account current vital sign values.\textsuperscript{18} In clinical practice, actualizing a clinician’s predictions into shared action and plans among the care team can be a significant challenge, in part due to the difficulties experts encounter in articulating these skilled judgements and the subsequent categorization of these predictions as anecdotal in nature lacking reliability and validity. Perhaps complicating the understanding of these predictions, is the notion that “it is not that proficient nurses have internalized the rules and formulas learned during the earlier stages of skill acquisition; they are no longer using rules and formulas to guide their practice. They are now using past concrete experiences much like the researcher uses paradigms.”\textsuperscript{19}

While we acknowledge the potential for bias and flawed intuition, we posit that the clinical expert intuition paradigm holds great potential as a novel and long ignored complementary factor in predictive modeling for current rule-based EWS models.\textsuperscript{20,21} Moreover, if the bias is systematic it can be computationally measured and leveraged for information.\textsuperscript{22–25}

EHRs contain large amounts of data that are never utilized or reused for clinical or analytics purposes. EHR log files contain a vast amount of data, some of which are rarely used yet can be extracted into computable metadata patterns that model healthcare processes.\textsuperscript{22,23,26} For example, these data provide the ability to measure patterns in order placement, medication use, note frequency, and start time for discharge orders. As part of a larger National Institute of Nursing Research (NINR) funded study – the CONCERN study - our team is mining nursing notes and flowsheets for clinical data that signals nurses’ concern. Our team uses EHR metadata to compute Healthcare Process Models for Clinical Concern (HPM-CC) that can be used to model proxy signals of clinical expert intuition that a patient is not doing well with enough lead time to intervene.\textsuperscript{20,21,24,25,27} HPM-CC are generated from perceptions, interpretations, and recordings entered by clinicians (e.g., nurses, physicians), and are based on clinician decisions to observe and enter data in the EHR.\textsuperscript{20,21,24,27,28} These types of EHR utilization patterns may be misinterpreted or discarded as biased for research, but these data and their biases are rich in information that can be used to understand clinical care and may enhance predictive power.\textsuperscript{22,23} For example, we have modeled the magnitude of these documentation concern patterns in relation to the clinical outcomes of in-hospital mortality and cardiac arrest.\textsuperscript{20,21} Similar analyses by other groups have shown laboratory test timing predicts mortality.\textsuperscript{29}

What mechanisms of action might explain why HPM-CC metadata patterns are a signal of clinical concern? Clinicians rely on situational awareness to continuously prioritize – or triage - care for the patients that they are most worried or concerned about. Clinical situational awareness is gathered from objective and subjective clinical assessments and expert pattern matching. Expert pattern matching is based both on the clinicians’ prior experiences with similar patients, and comparisons to the individual patient’s baseline health status. Although clinical nurses, like all experts, struggle to articulate their clinical intuition, 37 signs and symptoms categorized into 10 general indicators have been shown to underlie nurses’ worry or concern, or intuitive knowing related to patient deterioration in the hospital setting.\textsuperscript{30} While several indicators relate to vital signs, such as change in breathing, change in circulation, rigors and change in mentation, these physiological changes also may precede changes in vital signs.\textsuperscript{30} The remaining signs and symptoms are not related to measurement of vital signs: agitation, pain, unexpected trajectory, and patient indicates feeling unwell.\textsuperscript{30} Nurses engage in two activities as core parts of nursing practice that relate to expert pattern matching that result in EHR data with these HPM-CC signals. First, nurses increase their frequency of surveillance – and subsequently frequency of documentation – for patients that fit a concerning pattern.\textsuperscript{21} Second, nurses synthesize their subjective and objective clinical assessments - otherwise buried in structured flowsheet fields - within their narrative notes and short comments associated with specific flowsheet values.\textsuperscript{21,28} In other words, the act of documenting a free-text comment in a flowsheet row provides information that the nurse likely determined an event or observation was clinically significant enough to record, since these types of data are not required data entries. Predictive modeling that is focused only on the values of structured data in EHRs will miss the healthcare processes and nursing interventions that are activated far before a patient’s vital signs are abnormal.

While nurses’ subjective feeling of worry or concern is a valuable indicator before vital signs have even changed\textsuperscript{30,31} subjective nurse observations, such as “the patient does not look well”, by their very nature do not explain what is wrong or the rationale for why a patient does not look well. This is consistent with Benner’s characterization of clinical
expert intuition being challenging to articulate.\textsuperscript{16,19} HPM-CCs that identify when nurses make these statements, their associations with other signs and symptoms, and their ability to predict outcomes can lead to a shift in how we understand and leverage clinical observational skills and clinician entered data within a patient’s chart.\textsuperscript{24,27}

In this study we evaluate the performance of a broadly used and well validated EWS – the Modified Early Warning System (MEWS)\textsuperscript{32} – compared to what we call the MEWS-CC (MEWS-Clinical Concern) model. MEWS-CC is a novel early warning system which incorporates both the patient’s physiological data from the MEWS score as well as three HPM-CC proxy signals of a nurses’ concern identified from the CONCERN study (1. frequency of respiratory rate assessment, 2. frequency of flowsheet comment entered and associated with temperature value, 3. frequency of comment entered and associated with blood saturation) (see Table 1). In our prior analyses our team identified the significance of Sp02 comments for their ability to discriminate related to outcomes of mortality and cardiac arrest.\textsuperscript{20,21} We identified that Sp02 comments were significantly more likely to be associated with a normal Sp02 measurement, while comments for blood pressure and heart rate were more likely to be associated with an abnormal heart rate or blood pressure measurement.\textsuperscript{21,28} These Sp02 comments are used by nurses to highlight when a patient’s supplemental oxygen requirements are increasing; typically, in these instances a patient’s Sp02 reading remains normal because the nurse continues to titrate the supplemental oxygen in order to keep the Sp02 reading normal for the patient. Our prior work described this phenomenon and the nurse’s intention to highlight the relationship between Sp02 and supplemental oxygen - which EHR flowsheets do a poor job of visualizing – to the physician as an indicator of deteriorating status. Additional prior work from our team also identified that the frequency of respiratory rate and temperature comments have similar significant signals associated with mortality.

MEWS, which is based solely on a patient’s physiological data, is one of the most widely used and validated EWS models.\textsuperscript{32} Given the extent of its use in practice, and as a comparator to new EWS in other literature, was the most appropriate comparator for our study. We believe that this approach will demonstrate the added benefit of modeling predictive HPM-CCs compared to the MEWS criteria alone for the combined event of (whichever occurs first): transferred to ICU, transferred to SDU, rapid response, cardiac arrest or expiration\textsuperscript{4}, as well as time to first event.

Our analysis includes a specific focus on time from first alert trigger to event by simulating a real-time prospective analytical approach, accounting for the temporal signal of each data point. Therefore, we are pursuing a shift in the goal of predictive modeling. Rather than identifying when the patient’s values indicate they are already in a risky state, our team is focused on identifying when clinicians are concerned that a patient may be entering a risky state. The difference – we believe – is affording greater lead time (trigger to event time) to effectively intervene.

\textbf{Methods}

\textit{Inclusion criteria and data collection}

Flowsheet data was queried from two separate campuses of the same hospital in Northern Manhattan, a 745-bed adult facility in an academic medical center and a 300-bed community hospital, over a two-year period (January 2015 – December 2016). To minimize variation of the patient population three types of study units were identified, general medical or surgical acute care units (ACU) (n=12), intermediate care or step-down units (SDU) (n=4) and critical or intensive care units (ICU) (n=9). We collected data for patients that met our inclusion criteria of spending more than 24 hours on a study unit and excluded those patients that were less than 18 years of age on day of admission, patients who eventually had hospice or palliative care orders written (either upon admission or at any point during their hospital visit), and patients whose hospital stay exceeded 60 days.\textsuperscript{32} We captured five physiological parameters (heart rate, blood pressure, respiratory rate, temperature and level of consciousness) and an indicator of whether a flowsheet comment was written. The content of the comment was not used in this analysis. For each flowsheet entry we also captured the patient and visit identifiers and the recorded time. Patient visit admission, discharge, and transfer data was also queried so that patient locations could be associated to each entry in the flowsheet and patient movement within the hospital could be tracked. Patients and visits were identified by unique database generated identifiers.

Order and note data were also queried to aide in the cleaning and validation of the cohort of patients experiencing our primary outcomes. The order name and time the order was written was used to determine patients with hospice and palliative care orders. Note types, time of note creation, and structured fields within notes were queried to cross validate with discharge information the cohort of patients that expired or experienced a rapid response or cardiac arrest. The structured time field within these notes aided in the validation of times of events from other data sources.

\textit{Data cleaning and validation}
Outcomes data of patient expiration, rapid response and cardiac arrest were cleaned as mentioned above by cross validating the cohort and times of events between different data sources and finally performing a chart review on 20 selected patients. To determine the transfer information of each patient, we first looked at the distinct locations that a patient had any flowsheet entry on. We then used the earliest time that a flowsheet entry occurred on each unit to look up in the admission-discharge-transfer (ADT) table when the patient was transferred to this location. This enabled us to create a cleaned cohort of patients that we knew first spent time on an ACU before an ICU or SDU (by way of flowsheet entries) and the times that the patient transferred (by way of ADT entries). Ignoring this method and only focusing on ADT information resulted in patients that had transfer information but were never actually transferred. This method was validated through patient chart review.

Vital sign information was stripped of extraneous text so that they included only numbers consistent with possible physiological criteria for each measurement. To determine the level of consciousness for the MEWS score, one nursing expert mapped the top 95% of distinct structured and free text entries to the concepts ‘Alert’, ‘Reacting to Voice’, ‘Reacting to Pain’, and ‘Unresponsive’. These mappings were independently validated by a second nursing expert.

<table>
<thead>
<tr>
<th>EWS Models</th>
<th>Clinical Features</th>
<th>MEWS</th>
<th>MEWS-CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score = 3</td>
<td>Score = 2</td>
<td>Score = 1</td>
<td>Score = 0</td>
</tr>
<tr>
<td>Systolic Blood Pressure</td>
<td>(&lt;) 70mmHg</td>
<td>71-80 mmHg</td>
<td>81-100 mmHg</td>
</tr>
<tr>
<td>Heart Rate</td>
<td>(&lt;) 40bpm</td>
<td>41-50bpm</td>
<td>51-100bpm</td>
</tr>
<tr>
<td>Respiratory Rate</td>
<td>(&lt;) 9/min</td>
<td>(-)</td>
<td>9-14/min</td>
</tr>
<tr>
<td>Temperature</td>
<td>(&lt;) 35(^\circ)C</td>
<td>(-)</td>
<td>35-38.4(^\circ)C</td>
</tr>
<tr>
<td>Level of Consciousness</td>
<td>(-)</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>Clinical Concern</td>
<td># of Respiratory Rate Entries in past 24 hrs</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td># of SpO2 Comments</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td># of Temperature Comments</td>
<td>(-)</td>
<td>(-)</td>
</tr>
</tbody>
</table>

Table 1. MEWS criteria (blue), clinical concern criteria (yellow), MEWS-CC (MEWS-Clinical Concern) criteria (green)

Data analysis

Our data set was reduced to only include those visits that began on one of our 12 acute care units, as opposed to patients that were in an ICU or SDU prior to their ACU stay. We felt that this more accurately defined the visits that we were considering as unanticipated transfer to ICU or SDU. Upon review of the data and chart reviews, we did not feel that the hospital trajectory of a patient admitted to an ICU unit, transferred to an ACU and then transferred back to the ICU should be considered the same as a patient who began on an ACU and then transferred to the ICU.

The endpoints for our data set were discharged alive (no outcome) and the combined event of transferred to ICU, transferred to SDU, rapid response, cardiac arrest or expiration (primary outcome). The time for the primary outcome was the first time at least one of the sub-outcomes occurred.

The MEWS model and the MEWS-CC models (TABLE 1) were implemented in our sample data. The three additional clinician documentation pattern data points from the CONCERN study were: the number of respiratory rates entered.
(not considering the value, just that a respiration rate was entered), the number of SpO2 comments written, and the number of temperature comments written. These variables were selected based on our team’s preliminary analysis focused on patient mortality which indicates a strong signal using longitudinal logistic and hazard regressions with time-varying covariates. The scoring criteria displayed in Table 1 for each of these variables were naively selected based on inspection of the probability mass functions (pmf) for the number of respirations entered, SpO2 comments, and temperature comments. An estimate of the pmf for number of respirations entered in a 24 hour period was generated via a histogram, for patients admitted on an equivalent, but not included, ACU. The estimate showed a mode at 5 and approximately 50% of the mass to the left of 6, approximately 80% to the left of 11, and approximately 98% to the left of 21. For temperature and SpO2 comments, thresholds 1 and 2 corresponded to approximately the 98th and 99th percentiles, respectively. It will be the purpose of future work to analytically derive optimal thresholds for all of our covariates, however this work shows that positive results can be obtained by including crude estimates for clinician documentation frequencies.

As with many EWS models, the scoring system displayed in Table 1 is used to calculate a total score for which there is a threshold chosen that any patient with a score above that threshold is considered at risk and any patient below that threshold is considered low risk. MEWS, as a broadly used and validated score, has a widely accepted threshold of > 4 which we used for the MEWS in our analysis. However, we also ran the MEWS analysis for additional thresholds (0 to >9) to confirm any improved performance at other thresholds in our data set. Give that our new model MEWS-CC does not have established thresholds we computed performance measures for a series of different thresholds in order to understand which resulted in the best performance.

While our data set is retrospective, these two models were implemented using a simulated real-time approach – at each data point both models would update their scores as though a clinician had just entered this new data point in real time. This was done to provide an accurate baseline for our comparator, MEWS, as well as output an accurate time that both models would send alarms if implemented in real-time today. Since we know each patient visits’ endpoint, we could then compute time of the alarm to the time of the endpoint.

Finally, we ran the MEWS model alone and compared it to the MEWS-CC model to determine the change in predictive power. We approached the evaluation of the EWS models based on a previous study conducted by Umstead et al. 201633, since this approach includes both evaluation of predictive accuracy and evaluation of time from alert trigger to event (lead time). Based on Umscheid et al. 2016 we identified the following four metrics to describe differences in performance between the two models33 for the thresholds described above: 1. Area Under the Receiver Operating Curve (AUC); 2. Positive Predictive Value (PPV); 3. Negative Predictive Value (NPV); 4. Lead time i.e. how early the model was able to identify the adverse event before it happened (evaluation of time from alert trigger to event).

Results

A total of 44,581 unique patients (64,822 hospital admission encounters) were included in our initial data set. After applying inclusion and exclusion criteria our final data set was comprised of 28,394 unique patients (38,885 encounters). (See Table 2.)

Table 3 provides the performance metrics for each model, including 4 different thresholds of scores for the MEWS-CC model. The MEWS model, using the standard threshold of 4, had the following performance in our data set: PPV = 0.2642, NPV = 0.9247 with a median of 13 hours from trigger to event. The MEWS-CC threshold of 5 was most similar in performance to the MEWS with: PPV = 0.2348 and NPV = 0.926. The meaning of the false positive rate in this is complex. It describes the number of patients that were predicted to have one of our outcomes (transferred to ICU, transferred to SDU, rapid response, cardiac arrest or expiration) but did not. While the alert trigger was a false positive, it is possible that these patients may have had indicators of concern and been in a risky state that was mitigated by clinical intervention. Therefore, we interpret the true positive rates in Table 3 as at minimum the patients who experienced an outcome and we should be concern about.

Table 3 also displays the lead times for each model and threshold to allow for evaluation of the amount of time a clinician is afforded to intervene for an at risk patient after the EWS alert triggers. The lead time of the MEWS was 13 hours. The lead time for the MEWS-CC model was 39 hours (threshold > 3), 32.5 hours (threshold > 4), 24 hours (threshold > 5), and 18 hours (threshold > 6). Therefore, using a threshold of > 3 for the MEWS-CC model, there was an additional 26 hours, or greater than one full day, for clinicians to act and intervene on a patient with an alerted risky state. At the threshold of > 5 for the MEWS-CC model there was an additional 11 hours of lead time, equivalent to.
almost 1 nursing shift, for clinicians to act and intervene. MEWS was also run for additional thresholds and each time the two models had comparable performance, the MEWS-CC had a greater lead time. The MEWS-CC threshold of >3 contributed to a much higher false positive rate than MEWS-CC threshold of >5. The complexity of this false positive rate is assumed to include patients that were never at risk, but also patients that were at risk but did not experience a bad outcome due to early identification and clinical intervention or other unknown reasons.

Figure 1 displays the Receiver Operating Curves (ROCs) for MEWS and MEWS-CC Models and calculation of the AUC. The AUC for the MEWS model was 0.61. The AUC for the MEWS-CC model was 0.65. There was similar performance of AUC between the MEWS and MEWS-CC at a threshold of >5.

<table>
<thead>
<tr>
<th>Table 2. Characteristics of study data set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Admissions</strong></td>
</tr>
<tr>
<td>Patient Hospital Encounters, n</td>
</tr>
<tr>
<td>Unique Patients, n</td>
</tr>
<tr>
<td>Age at admission, years old</td>
</tr>
<tr>
<td>Male, n (%)</td>
</tr>
<tr>
<td>First adverse event in episode</td>
</tr>
<tr>
<td>None, n (%)</td>
</tr>
<tr>
<td>Death, n (%)</td>
</tr>
<tr>
<td>Cardiac arrest or Rapid Response, n (%)</td>
</tr>
<tr>
<td>Transferred to ICU, n (%)</td>
</tr>
<tr>
<td>Transferred to SDU, n (%)</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
</tr>
<tr>
<td>Number of observations, n (%)</td>
</tr>
<tr>
<td>Vital signs observations</td>
</tr>
<tr>
<td>Pulse rate observations, n (%)</td>
</tr>
<tr>
<td>Respiratory Rate observations, n (%)</td>
</tr>
<tr>
<td>Temperature observations, n (%)</td>
</tr>
<tr>
<td>Blood Pressure Systolic observations, n</td>
</tr>
<tr>
<td>(%)</td>
</tr>
<tr>
<td>Conscious Level, n (%)</td>
</tr>
<tr>
<td>Alert n = 346; Reacting to Voice n = 18; Unresponsive n= 6</td>
</tr>
<tr>
<td>Respiratory Rate, n (%)</td>
</tr>
<tr>
<td>SpO2 Comment, n (%)</td>
</tr>
<tr>
<td>Temperature Comment, n (%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Performance Metrics Reported per Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
</tr>
<tr>
<td><strong>Score Threshold</strong></td>
</tr>
<tr>
<td>MEWS 4</td>
</tr>
<tr>
<td>MEWS-CC 3</td>
</tr>
<tr>
<td>MEWS-CC 4</td>
</tr>
<tr>
<td>MEWS-CC 5</td>
</tr>
<tr>
<td>MEWS-CC 6</td>
</tr>
<tr>
<td><strong>True Positives</strong></td>
</tr>
<tr>
<td>MEWS 429</td>
</tr>
<tr>
<td>MEWS-CC 1451</td>
</tr>
<tr>
<td>MEWS-CC 912</td>
</tr>
<tr>
<td>MEWS-CC 522</td>
</tr>
<tr>
<td>MEWS-CC 289</td>
</tr>
<tr>
<td><strong>False Positives</strong></td>
</tr>
<tr>
<td>MEWS 1195</td>
</tr>
<tr>
<td>MEWS-CC 7761</td>
</tr>
<tr>
<td>MEWS-CC 3689</td>
</tr>
<tr>
<td>MEWS-CC 1701</td>
</tr>
<tr>
<td>MEWS-CC 771</td>
</tr>
<tr>
<td><strong>True Negatives</strong></td>
</tr>
<tr>
<td>MEWS 34455</td>
</tr>
<tr>
<td>MEWS-CC 27889</td>
</tr>
<tr>
<td>MEWS-CC 31961</td>
</tr>
<tr>
<td>MEWS-CC 33949</td>
</tr>
<tr>
<td>MEWS-CC 34879</td>
</tr>
<tr>
<td><strong>False Negatives</strong></td>
</tr>
<tr>
<td>MEWS 2806</td>
</tr>
<tr>
<td>MEWS-CC 1784</td>
</tr>
<tr>
<td>MEWS-CC 2323</td>
</tr>
<tr>
<td>MEWS-CC 2713</td>
</tr>
<tr>
<td>MEWS-CC 2946</td>
</tr>
<tr>
<td><strong>PPV</strong></td>
</tr>
<tr>
<td>MEWS 0.2642</td>
</tr>
<tr>
<td>MEWS-CC 0.1575</td>
</tr>
<tr>
<td>MEWS-CC 0.1982</td>
</tr>
<tr>
<td>MEWS-CC 0.2348</td>
</tr>
<tr>
<td>MEWS-CC 0.2726</td>
</tr>
<tr>
<td><strong>NPV</strong></td>
</tr>
<tr>
<td>MEWS 0.9247</td>
</tr>
<tr>
<td>MEWS-CC 0.9399</td>
</tr>
<tr>
<td>MEWS-CC 0.9322</td>
</tr>
<tr>
<td>MEWS-CC 0.9221</td>
</tr>
<tr>
<td><strong>Median Hours Lead Time</strong></td>
</tr>
<tr>
<td>MEWS 13</td>
</tr>
<tr>
<td>MEWS-CC 39</td>
</tr>
<tr>
<td>MEWS-CC 32.5</td>
</tr>
<tr>
<td>MEWS-CC 24</td>
</tr>
<tr>
<td>MEWS-CC 18</td>
</tr>
</tbody>
</table>
Discussion

Our implementation of a simple rule, based on frequencies of one vital sign entry and two vital sign comments, resulted in a model with equivalent performance to the MEWS that triggered alarms 5-26 hours (depending on selected threshold) before the first MEWS alarm. We emphasize that this approach was identified without using machine learning, but rather based on nursing subject matter expert knowledge of documentation and practice patterns for selection of variables that are a signal of nursing expert concern about a deteriorating patient. The ability to trigger an alert a full clinical shift (12 hours) or day earlier (24 hours) is clinically significant. For example, the clinical priority in order to prevent death for sepsis patients is early intervention which is dependent on early identification.\(^{34,35}\) It is plausible that nurses’ temperature comment documentation relates to the prominent campaigns in recent years to increase awareness and identification of sepsis.\(^{36,37}\)

Other studies have also identified EWS models with equivalent or slightly better performance than MEWS, such as Churpek et al. who used a machine learning approach, but - like many studies - did not report lead time.\(^{15}\) Importantly, we note that our team is approaching EWS modeling from a different paradigm of both data types (metadata patterns) and data temporality (simulated real-time prospective analysis). In our interpretation, a lack of reported lead time is ultimately incomplete for an EWS given that from a clinical perspective, slightly better accuracy of prediction without improvement in lead time offers limited opportunity to change patient treatment plan and trajectory. However, it is possible that combining newer EWS models that have improved accuracy, such as Churpek et al. (or other models derived from a machine learning approach), with our HPM-CC variables as used in MEWS-CC model could result in better performance and lead time that is both statistically and clinically significant. Further, our HPM-CC modeling approach is not a black box approach. In addition to validation of predictive power, our team is engaging in usability studies to understand the best approach to communicate predictions that are based on EHR metadata to busy clinicians to improve understanding of patient risky states and drive appropriate follow-up.

The complexity of the meaning of the false positive rates is critical to understand and account for given our aim to predict patients that are entering a risky state based on clinician concern. Further research is needed to understand if there were patients in our false positive group who avoided a bad outcome due to clinician identification of concern and early intervention - and if so it would appear that our data accurately detected that concern, but our study was not designed to detect early intervention. These false positive complexities are true of the MEWS group as well. Future
work should focus on discriminating between these types of clinical process characteristics in EWS studies. Similarly, other EWS studies have discussed differences between the choice of study endpoints in light of the fact that EWS impact on clinical outcomes is controversial.

Our analysis also included one difference in criteria that was not used in the published MEWS validation studies: patients that were admitted first to an ICU or SDU are excluded. This is a possible explanation of why we found a slightly lower MEWS AUC than reported elsewhere. In other words, patients that were admitted to the ICU and then transfer to ACU may be more likely to transfer back to ICU then a patient that was admitted first to an ACU. Both cohorts should be studied in future work. In the future, it would also be interesting to consider adding into an early warning system the indicator if a patient has ever been to the ICU and measure any EWS performance changes.

Our findings raise the question: if the documented nursing practice patterns - defined by our team as a proxy of nurses’ concern - are statistically significant early predictors why are these patients not adequately receiving early intervention once the nurse that documented those data is concerned? Clinical care teams and clinical workflows in the inpatient setting have complexities that have been shown to contribute to suboptimal communication paths and the efficient achievement of shared situational awareness among the members of the care team\textsuperscript{28,36,38–40} As discussed in the introduction, nurses’ expert concern is not always able to be well articulated with references to physiological changes to back up that concern, as typically desired by members of the medical team. This lack of a shared mental model of a patient’s state, combined with competing needs and priorities of other patients, contributes to delays in a shared situational awareness of a patient’s decline and delays in interventional care and procedures to save that patient. This is a challenging team situation in which the physician is typically not physically present at the patient’s bedside and able to observe the same subtle changes in the patient's condition that the nurse is responding to in our data. EHRs can help fill gaps in shared situational awareness for clinicians who are remote - not on site - but only for data types that are clinically well understood and recorded in the EHR. The metadata we processed in our study are not data types that are taught, referenced, or used in the course of clinical care. However, as mentioned, our team is investigating how these significant data points can be translated into an interpretable score by all members of the care team in order to increase shared situational awareness, even for clinicians that are viewing data remotely, to enable early intervention.

Our identification of respiratory rate as a significant signal for patient deterioration aligns with prior work describing oxygenation status and change in respiratory rate as one of the most common physiological criteria used to trigger a RRT\textsuperscript{12–14}, as well as Churpek et al., who identified respiratory rate as a predictive feature with significant vital sign interactions.\textsuperscript{15} From a clinical process perspective, respiratory rate is different from other vital signs in that, with the exception of ventilated ICU patients, it is typically manually measured. The nurse counts a patient’s respirations for a full minute. All other vital signs (blood pressure, heart rate, temperature) and related physiological measures (SpO\textsubscript{2}) are measured using a machine. However, in reality the manual process of counting a patient’s respirations for a full minute is typically not done in practice because an experienced nurse (and even most novice nurses) can typically identify a patient with abnormally high or low respirations upon initial visual observation. This practice results in nurses documenting respiratory rates less frequently in practice than other types of vital signs; we posit that beyond recording that a patient’s respirations were normal occasionally throughout a nursing shift, documentation of this vital sign may only occur when a patient’s respiratory rate is observed to be abnormal. Conversely, other vital signs require a specific machine measurement, rather than an “eyeball visual observation” and in clinical practice it is not common to take a machine measurement without recording it in the chart. However, an “eyeball visual observation” that a patient’s breathing looks ok – while likely performing other patient care activities - may not always be recorded in the chart. This difference in how, when, and why a physiological measure is performed and recorded in the EHR is important to understand when analyzing flowsheet data for EWS modeling because it impacts the information signals, including frequency of measurement and subsequently missing data. We also point to a potentially similar pattern of missing data in the clinical observation level of consciousness. It is possible that nurses document this selectively for patients for which it is more clinically relevant due to disease process (neurological) or clinical setting (ICUs). As stated previously, our past work identified the significance of SpO\textsubscript{2} comments for their ability to discriminate related to outcomes of mortality and cardiac arrest and that a comment is more likely to be associated to an abnormal SpO\textsubscript{2} value\textsuperscript{20,21} Our findings indicate that there may be similar clinical processes that factor into a nurse deciding to enter a respiratory rate value, SpO\textsubscript{2} comment, and perhaps temperature comments as well.

EHRs have been criticized for leading to information overload at the point of care. Yet, the vast amount of EHR information available can be extracted into computable metadata patterns that model healthcare processes and when synthesized can provide insights into clinical care patterns and associated outcomes. We believe that surfacing select patterns at the point of care could be a potential approach to overcoming information overload and supporting clinical
expert decision making by the entire care team. The findings from this study indicate the potential of HPM-CCs to enhance the lead time for triggering EWS and providing care teams with greater insight into patient’s risky states and ability to effectively intervene.

**Limitations**

The findings presented in this study are from two hospital sites part of one hospital system and require validation in other hospital data sets. To increase generalizability of our findings we used a well-validated EWS – the MEWS. We found overall similar performance of the MEWS in our data set as reported elsewhere in the literature, lending confidence in the data used in this study. It is important to note that the three selected HPM-CC variables are based on our team’s work to date which is not exhaustive of all nursing concern signals and additional significant variables may be identified from our data set as our work continues. We therefore position this evaluation as a use case to demonstrate the potential methods for enhancing EWS, rather than a validated and conclusive finding. Finally, the high level of missing data for the level of consciousness MEWS feature may have influenced the performance of the model, yet this is an evaluation of performance in using real clinical data, and handling missing data could be considered an inherent part of model performance in the clinical setting.

**Conclusion**

Unfortunately, there remain high rates of preventable deterioration in the hospital setting.2-4 Timely identification of patient deterioration is a modifiable factor that can change the trajectory of a patient’s status and outcome. We identified three variables from HPM-CC metadata patterns of nursing documentation as proxies of nurses’ concern that were able to increase the time from EWS trigger to event for patients in our data set. This is a noteworthy finding with direct links to potential clinical interventions, such as clinical decision support alerting and care team dashboards displaying patient risk status. Future work should confirm the best HPM-CC factors and thresholds to combine with EWS and validate these predictive models in at other sites.

**Funding Acknowledgements:** This study was funded by NINR 1R01NR016941-01, Communicating Narrative Concerns Entered by RNs (CONCERN): Clinical Decision Support Communication for Risky Patient States and the CRICO Risk Management Organization. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

**References**

A Sociotechnical Multiple Perspectives Approach to the Use of Medical Scribes: A Deeper Dive into the Scribe-Provider Interaction


¹Oregon Health & Science University, Portland, OR USA

Abstract

Because of increased electronic health record use, many organizations are hiring medical scribes as a way to alleviate provider burnout and increase clinical efficiency. The providers and scribes have unique relationships and thus, this study’s purpose was to examine the scribe-provider interaction/relationship through the perspectives of scribes, providers, and administrators utilizing qualitative research techniques. Participants included 81 clinicians (30 providers, 27 scribes, and 24 administrators) across five sites. Analysis of the scribe-provider interaction data generated six subthemes: characteristics of an ideal scribe, characteristics of a good provider, provider variability, quality of the scribe-provider relationship, negative side of the scribe-provider relationship, and evaluation and supervision of scribes. Future research should focus on additional facets of the scribe-provider relationship including optimal ergonomic considerations to allow for scribes and providers to work together harmoniously.

Introduction

The use of electronic health records (EHRs) in the United States has increased dramatically over the past decade due to the Health Information Technology for Economic and Clinical Health (HITECH) Act¹. Subsequent to the adoption of EHRs, unintended consequences like over-documentation² have occurred due in large part to regulatory mandates of documentation and documentation system inefficiencies. One way that organizations address physician dissatisfaction associated with the EHR and documentation is by hiring medical scribes. Mishra, Kiang, and Grant define medical scribes as “paraprofessionals who transcribe clinic visit information into the EHRs in real time under physician supervision³(p1468).” Prior research has demonstrated that the use of medical scribes lowered the documentation time for physicians³ and increased the time and quality of provider and patient interactions³,⁴. According to one study, medical scribes also “reduce the [provider’s] burden of record keeping and order entry⁵(p277).” Despite the documented benefits, concerns have been identified such as functional creep of scribe duties and the stifling of further development of EHR usability⁶.

Few comprehensive studies have analyzed the interaction of providers and scribes within the context of the sociotechnical environment in which they operate. Yan and colleagues⁴ conducted a seminal study that used qualitative research methods to investigate the role of clinical scribes through the lenses of providers, scribes, and patients in primary care. They described three themes: documentation, patient care, and teamwork (the scribe-provider interaction). They found that from the documentation perspective, the scribes captured information in real time and increased the amount of perceived detail in the providers’ notes. From the patient perspective, the patients perceived that provider spent more time focused on them and less time focused on the EHR by utilizing a medical scribe. From the teamwork perspective, high levels of communication and trust were important characteristics for both scribes and providers. Scribes also valued providers who were willing to teach. While this study had interesting findings, it only investigated primary care settings.

Sattler and colleagues conducted a one-year qualitative longitudinal study investigating four primary care providers and two scribes at one family care practice to learn more about the scribe-provider workflow and relationship⁷. They sent out surveys to the providers after each day and ultimately generated 11 themes, “which were further categorized under four domains: clinical operations, joy of practice, quality of care, and patient experience⁷(p51).” Regarding clinical operations, it was revealed that scribes lowered the amount of time providers spent writing notes. Surrounding the joy of practice, one physician said it was, “So fun to have a ‘partner in crime’ and to spend the day together⁷(p52)!” Another provider said that, “It’s also important to note that it is really fun having [the scribe] to work with. She is a friend now and a colleague⁷(p53).” In relation to quality of care, the study demonstrated that the scribes learned provider preferences and included more details than the providers did in their notes. The results surrounding patient experience revealed that the providers were able to spend more quality time with their patients and notice nonverbal communication they could have missed without the scribe in the exam rooms. The study had important limitations,
including that it only investigated a single primary care clinic, had a small sample size (four providers and two scribes) and did not survey scribes.

Neither of these previously mentioned studies used a sociotechnical model as a framework for studying medical scribes. One of the more popular sociotechnical models is the eight dimensions of sociotechnical model for health information technology. This model shows how all eight dimensions work together to help understand complicated healthcare entities.

The goal of the overall study was to use a sociotechnical model as a framework for studying scribes and the EHR in urgent care, emergency departments, and primary care settings across five organizations. The primary goal of the present study was to further expand knowledge of scribe-provider interactions and describe this relationship dynamic in-depth by analyzing what we discovered through interviews and observations from providers, scribes, and administrators. The overall purpose of this work was to investigate the climate of medical scribes in various practice settings and use this information to develop a set of simulation activities based on identified and needed competencies.

**Methods**

**Settings:** We used purposive sampling and selected our sites based on (a) the diversity of geographical location (b) size of clinics/hospitals, (c) whether the hospital/clinic was academic or community based, (d) EHR vendor, and (e) whether the scribes were hired through an internal program or a scribe company.

From October 2017 through January 2019, we conducted five site visits across the United States to hospitals and clinics in multiple geographic settings. The sites included teaching hospitals, community health systems, and outpatient clinics representing a wide range of sizes. Most clinics/hospitals were using the Epic system. However, one had recently switched from Cerner to Epic and one used AllScripts. We were interested in studying various scribing models; therefore, we visited one site that used virtual medical assistants (MAs) and nurses as medical scribes. Two of our sites used a traditional third party scribe company to hire scribes, which entailed hiring pre-professional scribes that were typically pre-health students. One of our sites had a homegrown scribe program where the hospital created their own pre-professional scribe company to hire and train scribes; the scribes were typically pre-health students. One of our sites used a scribe company that was closely tied to the academic hospital to hire pre-professional scribes; these scribes were typically college students with pre-health backgrounds.

**Data gathering:** We used the Rapid Assessment Process (RAP) model to construct our methodology. RAP uses a rapid qualitative and ethnographic approach to aggregate and process data in a timely manner. RAP helps establish ethnographic validity by using triangulation. Triangulation is “the use of several different researchers, the use of multiple perspectives to interpret a single set of data, and the use of multiple methods to study a single problem.” During the site visits we triangulated by using different researchers and methodologies.

We used qualitative research methods including observations in the hospitals/clinics to collect our data where researchers created field notes which are “usually handwritten notes that are done as the data is being collected.” We also used semi-structured audio-recorded interviews, which allowed the participants to answer questions on our interview guide, but they could also expand on topics or discuss ideas that were not mentioned in the interview guide. The team included ten researchers with multidisciplinary backgrounds including clinical, informatics, scribe vendor industry, and methodological expertise. Two trained researchers worked in pairs conducting the interviews. One person was the lead interviewer ensuring all questions on the interview guide were answered while the assistant interviewer asked follow up questions and made sure that the recorders were functioning properly.

To gather a holistic view, we took a sociotechnical approach to study medical scribes based on the eight dimensions of sociotechnical model for health information technology. During interviews and observations, we looked at the (a) hardware and software computing infrastructure, (b) clinical content, (c) human computer interface, (d) people, (e) workflow and communication, (f) internal organizational policies, procedures, and culture, (g) external rules regulation, and pressures, and (h) system measurement and monitoring. We used the eight dimensions of sociotechnical model for health information technology to study medical scribes. We investigated hardware and software computing infrastructure during our observations where we watched providers and scribes interact with the EHR. Through observations and interviews, we learned more about clinical content by hearing what vocabulary the scribes could add to the note. We observed human computer interface by watching the scribes interact with the EHRs. For people, we interviewed them and learned how the EHR made them feel and what kind of training they were provided to learn how to navigate the EHR. For workflow and communication, it was pertinent that we observed the
scribes’ workflow and see how they communicate with their providers. For a couple of the site visits, we were able to look at different organizations’ policies and procedures when it came to what scribes can and cannot do during a patient visit. It was pertinent that we investigated external rules, regulations, and pressures as well; we did by looking at the Joint Commission’s policies on medical scribing. Lastly, we looked at system measurements and monitoring through interviews and observations; we would ask how the scribes were monitored; all the organizations handled monitoring differently. We gathered data until we reached saturation, meaning until we were seeing and hearing the same things repeatedly.

Analysis: Using NVivo 11, we coded our transcriptions and iteratively conducted a theme analysis using a grounded theory approach. We began our coding process by using a dyad system to conduct a line by line analysis of the field notes and transcripts to identify repeated concepts, which we turned into codes. Once we created the codebook, we were able to use that to code our field notes and transcripts. Most of the transcripts and field notes were coded using dyad pairs consisting of the research team members who all had different backgrounds and allowed for a multidisciplinary approach. The team met often to discuss results and report findings, which is a form of member checking and helps validate the results. We used the codes we found in the transcripts and field notes to develop the themes and subthemes. Overall, we had ten coders who helped code the transcripts and field notes.

The study was approved by the Institutional Review Board (IRB) at the Oregon Health & Science University. IRB review and/or official determinations (if a site did not have an IRB, it had other mechanisms for reviewing our protocol) were also obtained from all sites.

Results

We visited five site visits across the United States with varying differences (see Table 1).

Table 1. Site visit demographics.
During our site visits, we interviewed a total of 81 clinicians (30 providers, 27 scribes, and 24 administrators) for a total of 47 hours of recorded audio. Physician interviews ranged from 9 minutes to 58 minutes; scribe interviews ranged from 11 minutes to 65 minutes; administrator interviews ranged from 21 minutes to 80 minutes. Most interviews took place in person, but one interview took place over the phone. It is important to note that nine of our interviews had only one interviewer instead of two due to members of the research team needing to be in other interviews or observing.

We observed at five sites, which included 12 clinics. Specifically, there were two Urgent Care Clinics, two Ear, Nose and Throat (ENT) Clinics, three Family Medicine Clinics, three Emergency Departments (EDs), one Pediatric Cardiology Department, and one Sports Medicine Department. We spent a total of 80 hours observing.

Through this research we discovered ten majors themes (see Figure 1).

**Figure 1.** Themes of sociotechnical research on medical scribes.

We have previously described ten themes emanating directly from our data, but this paper will focus on the scribe-provider interaction theme. The *scribe-provider interaction* theme is crucial in explaining how the provider and scribe work together effectively and how their relationship could improve. Through qualitative analysis, we found that this general theme included six subthemes: characteristics of an ideal scribe (knowledge/characteristics all exceptional scribes should possess), characteristics of a good provider (knowledge/characteristics all exceptional providers/scribe-users should possess), provider variability (how providers differ and the impact these differences have on scribes), quality of relationship (the benefits of the scribe and provider relationship), negative relationship (the negative aspects of the scribe and provider relationship), and evaluation and supervision of the scribes (what level of evaluation/supervision the scribes receive from the providers) (Figure 2).

**Characteristics of an Ideal Scribe**

During the interviews and observations, scribes, providers, and administrators offered descriptions of characteristics they felt scribes should exhibit and key knowledge a scribe should possess. One of the most commonly occurring characteristics was excellent communication. An ideal scribe can communicate openly with the provider and “ask a few questions” to clarify things. One scribe said that, “The more communication you have with your provider... the better it is for everyone.” The same scribe further went on to state that scribes are, “in a healthcare system and communication is the only way to get things done right. You can’t expect people just to read your mind.” With the unique dynamic between a provider and the scribe, the scribes need to be willing to communicate their needs. As a
provider explained, “one of the hallmarks of a great scribe too is speaking up for yourself. We are moving so fast in clinic that I feel like a good scribe will [ask questions or]...stop me so they can fill in the note.” Scribes also need to be able to handle criticism and feedback too. As one scribe stated, “Your provider is going to give you criticisms, and we talk about ways to kind of turn that into a positive rather than negative.”

It is important for scribes to be flexible/adaptable as well. Many of the scribes we interviewed and observed worked with multiple providers and in multiple clinics/specialties; they had to adapt quickly to the provider’s preferences and department differences. One administrator said that scribes “aren’t the one that’s leading the pace or direction of this investigation. [They] are along for the ride and sort of adjusting the entire time.” Because they have to adapt to situations quickly, being able to handle pressure is critical. In reference to one urgent care clinic, a scribe stated that, “getting flustered, easily, isn’t going to work well in this atmosphere.”

An ideal scribe should have some sort of medical/healthcare background (e.g., current pre-medical student, prior work as another healthcare professional [MA, CNA]). Due to the fast-paced environment of scribing, knowledge of medical terminology (including anatomy and physiology) was mentioned multiple times as something scribes need to know well. Otherwise, a scribe might be inefficient if they struggle with terminology.

Professionalism, not only in appearance, but also in interactions towards patients and providers is key for scribes to be successful. As one scribe mentioned, a scribe needs to be “a fly on the wall, because [patients] are going through hard things and it’s hard enough to talk about [it] to one person, let alone one person and their helper in the corner...so just being respectful.” A provider told us she “wanted the scribe to be a fly on the wall, to melt into the corner, to not be a part of the conversation, to be an observer.” One site uses remote scribes to achieve this idea of “fly on the wall.” The scribes talk to the providers through audio only, with no video recording, which allows for the patients and providers to talk freely without the scribe physically present.

Scribes need to be able to type quickly with accurate spelling and grammar, and possess strong technology/computer skills (e.g., knowing how to navigate EHR). Scribes have been called the provider’s, “peripheral brain” because, “The scribe is [their] go-between the medical information being recorded and the scribe is also interacting with the [provider in] a way so [they] can interact with the patient in the best way possible.”

A scribe needs to be intelligent and a “quick learner.” One provider mentioned that pre-professional scribes “are just sponges, and they are so smart and so good at learning.”

Lastly, scribes, especially pre-professional scribes, must have a passion for medicine. As one scribe noted, scribes need to “be interested in medicine... that doesn’t mean being pre-med or pre-PA or whatever it may be, just having a genuine interest in learning everything about medicine.” A few people we interviewed and observed stated that having

---

**Figure 2.** Subthemes of the scribe-provider interaction theme.
a passion for medicine ties in to motivation. One provider said, “I think a scribe who didn’t have an interest in medicine, I think, for obvious reasons, wouldn’t be motivated to learn the [medical] lingo necessarily.”

One scribe told us that her “motivation and interest in one day being a doctor” helps her be a great scribe. Another scribe said that the most important thing about being a scribe is that “you need to be motivated and actually enjoy medicine.” This scribe further went on to say that when he became a scribe he told himself that he wanted to become a PA “so everything that [he] was learning [he] was retaining.”

Characteristics of a Good Provider

Scribes, providers, and administrators often listed characteristics they felt providers/scribe users should have. Like an ideal scribe, a good provider is able to communicate effectively. One scribe mentioned that, “The more that a physician’s willing to communicate, the more smoothly that the day goes.” Another scribe noted, “It can get a little frustrating when they’re not communicating with you, or when they’re just doing their thing, and they pretend you’re not there.”

The scribes also want to work with providers who are willing to clarify exam findings and are clear about what they want in their note. One administrator told us that “the scribe’s role is purely… to do documentation under explicit direction and so an ideal provider is someone who realizes this and is very vocal and is able to think out loud.” A few scribes mentioned that they like it when a provider says the exam findings aloud during the exam, which allows them to be more efficient with their notes.

Staying organized is another key characteristic of a good provider; a scribe said that if the providers “are disorganized... it puts [them] behind because [they] cannot obtain the information that [the scribe] need[s] from them.”

One administrator said scribes “grow based on their feedback” from physicians, so scribes also want a provider who is willing to provide them with feedback, even if it is negative, so they can improve.

One of the most reoccurring characteristics of a good scribe user is one who is willing to teach. When talking about trying to find providers to use scribes, one administrator said that they “look for that doctor that has that teaching heart.” During an observation, one researcher noted that one provider was a “natural teacher, so he love[d] answering [the scribe’s questions] and explaining medicine to [the scribe].” A scribe stated that he thought the providers “[knew] scribing [was] not always the most exciting job. And so they [did] their best to help make it educational for [the scribes] and make it a good experience.”

Another scribe said that “being willing to teach your scribe [is important]. Like, we can only learn so much on our own. We can do all this research at home; but if you talk us through things, we’re not going to forget it, you know, when we’re seeing the flu or strep throat. And, you’re like, okay, this is what’s going to be. This is how we’re going to treat this. These are the generic symptoms for this. If they take the time, I feel like they put out what they put in. So, if they want to be our teacher, they’re going to get better charts out of it.”

A good provider must be willing “to give up a certain amount of control” when it comes to specific word usage in a note. One administrator said that, “A lot of [the providers] have a hard time [letting go of control of the note] and they need every word to be literally what they would have written.” They need to be able to “be comfortable with some degree of variation in the phrasing” of sentences.

One provider said that “some [providers he] knows are so meticulous and so detail oriented that everything has to be perfect... [he] thinks that people have to be flexible” when it comes to the note. The providers must review the scribes’ notes, but letting go of small structural changes or latitude changes in phraseology could allow the providers to review charts quickly.

Finally, a provider must have patience “because sometimes [the providers] forget that it is a human working the computer” and because “not every scribe is perfect on their first shift.” Scribes are an investment; as one provider said, “the more we invest in them the faster they learn and the better they get. So, it’s really on us to improve them.”

Provider Variability

Each provider is unique. As one researcher noted during observations, “the scribes usually said there were a variety of personalities and some were easier to deal with than others, but they like the variety.” This was the case for most
of the sites. At every site, we observed that scribes had profiles that were specific to each provider and included provider preferences, templates, smart phrases/macros, general information on the specific clinical workflow, and common terminology. It is key for the scribes to be able to adapt to the provider’s variability.

One scribe summarized this: “Every provider is different and knowing the preferences before you walk in, you will have a better shift. If you walk in not knowing who you are working with or the preferences, I think it is going to be a rough shift... I have to come in maybe 15 minutes early and set up more macros than I would normally, or ask them more questions or be ready to type paragraphs on paragraphs on the HPI because they are so in-depth, versus other ones that are more laid back.” Maintaining profiles with provider preferences was especially useful for scribes who worked across specialties with multiple providers throughout any given week. Because providers have different preferences in their notes, a few sites have tried to improve this process as much as possible by creating a standardized note template for scribes to use across all providers.

Quality of Scribe-Provider Relationship

The scribe and provider have a unique relationship; as summarized by a scribe, “there are very few people in the clinic, if anybody, who spends as much time with the provider as the scribe does.” Because of the amount of time spent together, setting up a functional relationship between the scribe and provider is critical. The providers want to be “able to bond with [their] scribes” and the scribes want that relationship too. The scribe and provider must have a good rapport and work together as a team, despite the power dynamic. An administrator stated, “of course, there’s a hierarchy here but also, [the providers] are relying on [the scribes].” Another provider stated, “It is crucial to that relationship that the physician view the scribe as someone who’s a key member of the team.” As one provider noted, having a stronger overall relationship with his scribes led to a higher “quality of note that results as a consequence.”

Trust is a key component to this relationship dynamic as well. One administrator summarized, “if you don’t have that trust set up, you are [out of luck]. Because that is the foundation of scribing.” Typically, the longer a scribe is with a provider, the higher the level of trust can develop. Almost like a dating website, matching the scribes and providers is a key component to forming a high quality relationship. Scribe companies must determine what the provider’s personality is as well as the department’s culture, and try to find a scribe who will mesh well with the entire healthcare team. When the scribe-provider relationship is well formed, one provider noted that “we leave happy [at the end of the day]. We leave with a smile on [our] face.”

An administrator provided an example of how one scribe had been with the same provider for the past five years and the provider “loves her. She’s part of the family. [The provider] is not going to quit. He’s going to work forever”. If a solid relationship formed, the scribe can often predict the provider’s next move and anticipate what questions the provider will ask their patients. One provider told us that the scribes “are not transcriptionists at all... they’re pulling out things and dropping down things, and managing things, and gathering information from other sources for us. They’re doing stuff that I don’t feel competent to do. And I always tell them at the end of the day, thanks... we both do what we do best, and we, both complement each other.”

Negative Side of Scribe-Provider Relationship

While there are many positives to the scribe and provider relationship, during our site visits we also noted potential negative issues. One issue that was identified in multiple sites was the lack of provider revisions of scribed documentation prior to sign off. One administrator stated that she “would guess a lot of the doctors, once they feel really comfortable with a scribe they do a quick review if they are being good [as in, being thorough]. But some of them may just sign and that’s where you get into the risk.” Another administrator mentioned that he “thinks that [providers] get kind of lax when they have had a scribe for a while... [and] read through [the note] but maybe just a lot quicker.”

A different administrator mentioned that she saw scribes “writing pretty junky notes... and the doctors were just signing them. Not reading, just sign, sign, sign, sign.” Another provider stated that “if I know my scribe is accurate almost all the time, then I can more quickly review the note and sign it off.” Most of the providers are aware of the importance of reviewing, though. As one provider stated, “it is our responsibility ultimately to edit that note.”

Despite efforts to match scribe and providers carefully, mismatches of clashing personalities sometimes occurred. For example, two administrators told us about a mismatch between the scribe and provider based on political preferences. One provider stated that, “Sometimes the scribe’s personality doesn’t work out and it does not mean they’re a bad scribe or there’s something done wrong.” Mismatches were also common because the provider and scribe had
communication issues. One scribe talked about how they “rely on [the provider] to communicate... ‘cause if they’re not then it’s not gonna be a good work environment for either of [them].”

Departmental mismatches occurred as well. As one provider stated, “There’s some scribes that’ll come to us from another department where it’s an endocrinology visit and that person spends an hour-and-a-half with each individual patient and they see ten patients or whatever a day. That person has a hard time getting fast at Urgent Care type visits because you’re seeing people fast.” Mismatches also occurred because the providers lacked patience with their scribes who were new/inexperienced or slower. A few of the organizations try to alleviate mismatches between the scribe and provider by moving scribes to different departments that best match their personality and workflow.

The provider’s overreliance on scribes is another negative aspect that can lead to loss of productivity. Providers can become over-dependent on their scribes to write notes and navigate the EHR. Another provider mentioned the cognitive impact scribes have. He told us that “when he writes or types [the notes] himself, it goes into his brain and memory better. He finds himself asking the scribe to look things up that normally he would have remembered [if he had written the note himself].”

Another common negative was failing to make the scribe feel like part of the clinical team. Despite the efforts, the scribes sometimes felt like they were at the “bottom of the totem pole” when it came to working with the providers. While scribes are supposed to be part of the team with the provider, sometimes the providers asked the scribes to do tasks beyond their scope and the scribes complied. One administrator said that he has seen smaller clinics where “the providers are having the scribes log in as them and sign orders.”

Boundary issues may also occur. One administrator said that “scribes are bright, enthusiastic young people and they form personal relationships with the department and the physicians. Sometimes there are boundary issues with scribes.” This administrator said she has seen “things like scribes housesitting for [providers], scribes running errands for [providers], scribes doing nice things for doctors in expectation that will give them a better shot at a good letter [of recommendation].” With the power dynamic, it is the organization and provider’s responsibility to be clear and set boundaries.

Evaluation and Supervision of Scribes

An overarching theme we observed was the lack of standardization of evaluation and supervision of scribes. The evaluation and supervision of scribes seemed to vary not only between sites but also between individual providers within the same organization. A few of the scribe-provider pairs reviewed the notes after each patient, others reviewed notes together at the end of the day, and others providers reviewed the scribe’s notes on their own time away from the scribe. Each provider varied in the amount of feedback offered to the scribe. Occasionally providers offered regular feedback, which seemed to be valued by the scribes. However, not all providers offered much feedback. When this lack of feedback occurred, a couple of the scribes said that they learned to look back at past notes and see where changes were made, then they could adjust their notes in the future for that provider and learn from their mistakes. The scribes, however, do not have the clinical/medical background that the providers have to understand why these changes were made.

Trust plays a large role in the time providers spend evaluating scribes. As the provider’s trust increases, the amount of time reviewing notes and evaluating scribe performance decreases. One provider told us that “you see, you know, 25-30 patients and that, you know, three or four shifts a week, I can’t pore through everything. So, I pretty much trust my scribe on the history, the physical, the review of systems.” A scribe mentioned that “over time, providers start to trust you and know what you’re writing. And, obviously, I didn’t go to med school for 12 years, so I could have some errors in them, in my notes that I guess could be reviewed a little bit more in-depth... I guess the provider just kind of assumes that you know what you’re writing.”

The processes for scribe evaluation varied between study sites as well. Several of the sites had an informal process for evaluating a scribe’s performance. Typically, in these cases, the scribe management team would observe the scribe and offer remedial training if a provider had complained. The evaluation was more reactionary than proactive. Other sites had a more proactive approach to evaluating scribes. These sites had quarterly/monthly formal audits that took place where the scribe management team would audit a random selection of the scribes’ notes using specific grading criteria (e.g., did the chart get closed in a timely manner, did the scribe add their attestation). The scribe management team allowed the providers to “get a chance to evaluate the quality of scribe documentation as a part of the audit process.”
One of the scribes who audits the organization’s charts said that she “sends surveys to the providers each month, and [they’ll] rate them on... overall attitude, efficiency in charting, and so that’s their opportunity to... voice their concerns.” At the end of the quarter/month, the scribes would receive a document “about what [they] did right...[and] what [they] did wrong.” One scribe said that these monthly audits were “the way that most people learn.”

Discussion

This study is unique in its analysis of the scribe-provider interaction, and identification of components that make for a strong scribe and provider relationship. While there were overarching themes, we learned that there is great variability in what it takes for the scribe and provider to have a successful relationship. Every site did things in a slightly different way from the others; there was no observed standardization in how the scribe-provider interaction took place. Seeing the extent of variability between individual provider’s preferences and how organizations addressed this issue was fascinating. Some standardized providers’ notes as much as possible while other organizations asked scribes to keep track of different provider preferences.

This study found similar results to Yan and colleagues⁴; we saw that trust and communication are important characteristics for both scribes and providers. Furthermore, we also saw that providers need to be willing to teach their scribes and scribes need to have the necessary medical knowledge to write a note. The current study also expands on Yan and colleagues’ research. Their research specifically targeted primary care; we included urgent care and emergency departments as well as primary care in our study to show a more global picture of the scribe-provider interaction. We also observed the camaraderie between the providers and scribes noted in Sattler and colleagues’ research⁷. Their research focused on primary care providers, where our study branched beyond primary care. They also had a small sample size that focused only on the physicians’ responses and we conducted interviews with 81 individuals (scribes, providers, and administrators).

The current study does have limitations. First, since we saw wide variation in scope and use of scribes in clinical practice during our site visit, we suspect that the results of this study cannot be generalized since we expect the degree of variability to be widespread in all areas of scribe use. Second, all of the sites we visited volunteered to participate in the study, which could skew data or allow for bias. It would be beneficial if future studies investigated what the best ergonomics are for making the scribe-provider interaction more efficient. Through observations, we noted that the layout of the exam rooms could sometimes hinder how well a scribe could document the patient-provider interaction and there is very little research on ergonomics in regards to the scribe-provider relationship. Finally, determining best practices for scribes and providers/scribe users is key for making the relationship function. Future research should determine the best practices necessary to optimize this domain.

Conclusion

If scribes are to efficiently and safely assist providers in generating documentation using the EHR, the relationship between scribe and provider pairs need to be a constructive one. The scribe and provider need to be matched well in order for this relationship to blossom. The providers must supervise and evaluate their scribes’ performance as well. It is just as important that providers learn to be good scribe users as it is that scribes learn to do their job well. Providers who use scribes need training about how to use scribes most effectively. An important aspect of using them effectively is to continuously and deliberately promote positive relationships with them.

Acknowledgements

This project was supported by AHRQR01HS025141. We would like to thank Jeremy Liu, Renee Kostrba, Peter Lundeen, and Marcia Sparling for their help at each site as well as the administrators, providers, and scribes for their participation in this project.

References

2. Bates DW, Landman AB. Use of medical scribes to reduce documentation burden: are they where we need to go with clinical documentation? JAMA Internal Medicine 2018;178(11):1472-73.
A Factored Generalized Additive Model for Clinical Decision Support in the Operating Room

Zhicheng Cui1, Bradley A Fritz, MD2, Christopher R King, MD PhD2, Michael S Avidan, MBBCh2, Yixin Chen, PhD1

1Department of Computer Science and Engineering, Washington University in St Louis, St Louis, MO; 2Department of Anesthesiology, Washington University in St Louis, St Louis, MO

Abstract

Logistic regression (LR) is widely used in clinical prediction because it is simple to deploy and easy to interpret. Nevertheless, being a linear model, LR has limited expressive capability and often has unsatisfactory performance. Generalized additive models (GAMs) extend the linear model with transformations of input features, though feature interaction is not allowed for all GAM variants. In this paper, we propose a factored generalized additive model (F-GAM) to preserve the model interpretability for targeted features while allowing a rich model for interaction with features fixed within the individual. We evaluate F-GAM on prediction of two targets, postoperative acute kidney injury and acute respiratory failure, from a single-center database. We find superior model performance of F-GAM in terms of AUPRC and AUROC compared to several other GAM implementations, random forests, support vector machine, and a deep neural network. We find that the model interpretability is good with results with high face validity.

Introduction

Patients undergoing surgery and anesthesia experience external stresses that place them at risk for numerous complications, including acute kidney injury and acute respiratory failure. One of the roles of the anesthesia clinician is to regulate the patient’s physiology to minimize these risks. Logistic regression-based models for predicting postoperative acute kidney injury1–3 and acute respiratory failure4–6 have been developed by multiple groups. Linear models offer a high degree of transparency regarding which features drive the output, but they are inherently limited in their flexibility, which limits their predictive accuracy. Various machine learning (ML) models have been proposed to solve these clinical prediction tasks with greater accuracy. Classifiers for acute kidney injury have been described in postoperative7 and non-surgical hospitalized patients8, 9. Although these ML models outperform logistic regression, they are not frequently used in clinical practice in part due to their lack of interpretability.

An interpretable model must provide predictions that are both accountable and actionable. An accountable model provides information about which features are contributing to the output prediction. This information can include feature importance and feature interactions. An actionable model provides guidance regarding how to modify the input features so that the post-intervention features will lead to the desired output.

Interpreting ML models is an extremely active field10–12. Incorporating interpretability constraints directly into the structure of the model and using post-hoc interpretation methods are two of the main directions13. Most existing work focuses on accountability. Che et al14 transfer the DNN’s knowledge to gradient boosting trees (GBT) using knowledge distillation and interpret feature importance through measures of variable importance designed for GBT. Another work15 visualizes the region of interest through class activation maps. Ge et al.16 feed features extracted from recurrent neural networks into a logistic regression model for prediction, where importance of the transformed features can be directly read off. Neural networks with attention-like mechanisms are also popular to visualize the features which contribute the most to a classifier for a particular case11. A smaller number of techniques address the actionability requirement. An early work17 proposed an integer linear programming method to extract actionable knowledge from a random forest. Gardner et al. proposed a label changing method by searching semantically meaningful changes to an image under its manifold space18. As far as we know, little work has been done on addressing accountability and actionability at the same time in the clinical area.

An extension to generalized linear models, generalized additive models (GAMs), can address accountability and actionability simultaneously. Examples includes LR, density based logistic regression19 (DLR), generalized additive
neural networks (GANN), and deep embedding logistic regression (DELR). LR assumes a fixed (up to a parameter) monotonic relationship between each feature and the outcome probability, limiting its flexibility and predictive performance. GAMs loosen these assumptions by inferring a transformation of the inputs with the full flexibility of non-parametric or parametric methods. For example, DLR transforms each feature through a kernel estimator, and our recently proposed DELR performs feature-wise nonlinear transformation using neural networks. GANN, which used a single hidden layer, can be treated as a special case of DELR, which used multi-layer DNNs. Despite the increased flexibility, with suitable constraints we can extract accountability and actionability from GAMs. Given an input example, GAMs allow us to calculate the contribution of each feature to that example’s predicted value. Feature contribution curves can be drawn to provide actionable directions on the optimal change and magnitude of improvement for each numeric feature. However, only when all the features are conditionally independent (given the label) can GAMs model the true distribution of data. Feature interactions are not allowed in GAMs, restricting their performance in dealing with complex datasets. In addition, GAMs have the undesirable property of treating static and time-varying features equally. For example, demographic characteristics such as age, gender, and height are not possible to change. On the other hand, it is possible to deliver interventions that modify a patient’s vital signs during surgery.

To address these problems, we propose a variation of GAMs that splits features into time-varying (or targeted) features and static features. F-GAM fits a context-based scaling for each time-varying factor based on the static factors, substantially increasing its flexibility compared to models which require the effect of a feature to be the same for all examples, but retains the ability to derive personalized feature-effect curves. F-GAM retains the full flexibility of a DNN for the effect of static features and DNN-based flexibility for the transformation of time-varying factors. We implement F-GAM as an end-to-end trained model with minimal hyper-parameters. In extreme cases where there are no static features available, F-GAM reduces to DELR. If there are no time-varying features, F-GAM becomes a DNN. We empirically validate the accuracy performance of F-GAM with existing ML models, including other GAMs and demonstrate the interpretability of F-GAM through a case study on predicting acute kidney injury.

**Background and Notation**

**Notation**

Operating room data contains both preoperative data such as demographic information and intraoperative data such as vital signs and medications administered. Given a patient \( i \), pre-op data \( x_i^S \in \mathcal{R}^{D_1} \) collected before the surgery are treated as static feature vectors while intra-op data represented as \( x_i^{TV} \in \mathcal{R}^{D_2} \) can be modified in real time. Together, we use \( x_i = [x_i^S, x_i^{TV}] \in \mathcal{R}^D \) to denote input features and \( y_i \in \{0, 1\} \) to represent the binary outcomes.

Our examples are binary classification, but the extension to multi-class classification is straightforward with a final softmax transformation and appropriate loss function.

**Generalized Additive Models**

A generalized additive model (GAM) is an ensemble of \( D \) univariate functions, where \( D \) is the number of features. We use \( x_j \) and \( y \) to denote the \( j \)th dimension of input \( x \) and class label, respectively. The output of each univariate function, denoted as \( f_i(x_j) \) is a real number. We can write the GAM structure as

\[
g(E(y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_D(x_D),
\]

where the function \( g \) is the link function, bounding the range of right hand side value of Eq.1, and \( E(y) \) is the expected value of the label conditional on \( x \). Constraints on \( f_k \) such as smoothness or degrees of freedom regularize the estimation problem to decrease out-of-sample loss. With a little abuse of notations, we use \( F(x) \) to denote \( E(y|x) \) throughout this paper for ease of presentation. By inverting the link function, the GAM has the form,

\[
F(x) = g^{-1}[\beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_D(x_D)],
\]

where the model output is controlled by the sum of each univariate function. GAM assumes all the features of input \( x_i \) are making contributions independently. Interpreting a GAM is straightforward as the marginal impact of a specific feature does not rely on the rest of features; we are able to know the importance of a feature by plotting its corresponding univariate function or calculating its variance over the sample. Actionable changes can be made based the shape of each \( f_k(x_k) \).
Figure 1: Overall architecture of F-GAM. Each circle denotes a scalar. Upper left part is feature mapping module. Every time-varying feature is fed to its own deep and narrow neural network (DNNN) separately. Weight learning module, which is shown in upper right part takes static features as input and calculates feature weights. Note that bias learning module is not plotted in this figure for simplicity.

Logistic regression is a special case of GAM by choosing logit function \( g(x) = \ln \frac{x}{1-x} \) as the link function and setting \( f_k(x_k) \) to be \( w_k x_k \) yielding

\[
F(x) = \sigma(w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D),
\]

(3)

where the sigmoid function \( \sigma(x) = \frac{1}{1+\exp(-x)} \) is the inverse form of the logit function. LR assumes a monotonic relationship between the final output \( F(x) \) and input features due to the linear function \( f_k \). However, this condition doesn’t hold in many cases, such as the relationship between ICU transfer rate and age.

**Methods - Model Algorithm**

In this section, we propose a factored generalized additive model (F-GAM) framework in which interactions between time-varying features and static features are allowed. The overall model has the form

\[
F(x) = \sigma \left[ \sum_{t=1}^{D_2} w_t(x^S) f_t(x^TV) + w_0(x^S) \right]
\]

(4)

In F-GAM, \( w_t \) is no longer a constant weight parameter, but the output of a DNN that accepts the static feature vector as input and estimates the weight of \( t \)th time-varying feature for each case. The feature-wise nonlinear transformation functions \( f_t, t = 1, 2, \ldots D_2 \) are jointly estimated. \( w_0 \) is a bias / intercept term that also depends only on the static features. In our operative examples, \( w_0 \) represents the estimate of risk before any intra-operative data becomes available as long as the input features have been appropriately centered.

F-GAM can be decomposed into four different modules: time-varying feature mapping module, feature weights learning module, bias term learning module and logistic/softmax regression module. We display the F-GAM architecture in Figure 1.

**Time-varying feature mapping module**

In traditional GAMS, the ability of the univariate function \( f_k \) to approximate the unknown transformation plays a crucial role in model performance. We choose to use deep and narrow neural networks (DNNN)\(^2\) for the nonlinear feature embedding. Being a universal approximator, a DNNN is able learn complex patterns automatically. The general tools for regularizing neural networks are immediately available to control overfitting without the difficult-to-understand
smoothness or degree-of-freedom constraints of other GAM transformations. Each time-varying feature is fed into a DNNN with distinct parameters; however, several hyperparameters (depth, width, dropout, training stopping time) are shared across $t$ to avoid having to search over a large hyperparameter space. The shared architectural parameters also tend to prevent over-fitting of just a few features (data not shown). A learnable look-up table (categorical embedding) is attached before a DNN for categorical features. In our examples, all time-varying features are quantitative or ordinal rather than categorical.

**Feature weights learning module**

Rather than applying fixed weights for the input features, we use nonlinear functions $w_t$ to adjust the feature weight dynamically. The nonlinear function should have the following two properties. First, the nonlinear function should not increase the number of parameters dramatically. Second, the nonlinear function should be able to handle both numerical features and categorical ones. Thus, we choose to use deep neural networks as the nonlinear functions. Rather than assigning each $w_t$ a standalone DNN as we did for $f_t$, all the weight-learning functions are estimated with a common DNN except the last layer. With this multi-task setup, we are able to exploit the shared structure of the data to reduce the effective number of parameters. Joint predictions of $w_t$ also allow the module to dynamically choose between potentially correlated $x_t^M$ to emphasize, meaning that $w_t$ represents both the relevance and precision of $f_t$ in the given context. For the second property, we use categorical embedding. When there are no static features, $w_t$ is a constant per time-varying feature and F-GAM reduces to DELR. That is to say, DELR is a special case of our model.

**Bias term learning module**

In order to increase expressiveness of the final model, we add a bias term based on the static features. Again, we use a DNN to model the bias term. This DNN is appended to the penultimate layer of the feature weights module to reduce redundancy. When there are no time-varying features, only the bias term controls the final output. In this case, F-GAM simplifies to a deep neural network.

**Logistic/softmax regression module**

With all the transformed features and weights ready, we apply the dot product operation to the time-varying feature mapping and the learned weights. After adding the bias term $w_0$, a sigmoid function $\sigma$ is used to model the positive rate given input data.

Our F-GAM is trained end-to-end by minimizing the cross entropy loss between true label distribution and prediction distribution. We also apply weight decay and an early stopping strategy to avoid over-fitting. The code is available at https://github.com/nostringattached/FGAM.

**Methods - Experiments**

**Data Sources**

Models were trained and validated using a dataset obtained from a single academic medical center (Barnes Jewish Hospital, St. Louis, Missouri). All adult patients who received surgery with anesthesia between June 2012 and August 2016 were eligible for inclusion. Due to the limited incidence of acute respiratory failure among patients who were not admitted to the intensive care unit (ICU) after surgery, prediction of this complication was limited to patients admitted to the ICU after surgery.

Acute kidney injury and acute respiratory failure were the two complications that were used as targets in the experimental models. Per Kidney Disease: Improving Global Outcomes (KDIGO) criteria, acute kidney injury was defined as an increase in the serum creatinine value by $>0.3$ mg/dL or $>50\%$ within 48 hours, compared to the preoperative value\(^23\). Acute kidney injury was undefined if the patient was receiving dialysis before surgery. The preoperative creatinine was the most recent value available before surgery, but no more than 30 days before surgery. Acute respiratory failure was defined as mechanical ventilation for $>48$ hours after surgery or reintubation within 48 hours. Acute respiratory failure was undefined if the patient was receiving mechanical ventilation before surgery, if the patient had a second surgery within 48 hours, or if the patient died within 48 hours.

Baseline demographic characteristics, comorbid health conditions, and preoperative laboratory values were retrieved
from the electronic medical record. The total doses of commonly used medications (including intravenous fluids, blood pressure-raising and -lowering agents, sedatives, pain medications, and nephrotoxic antibiotics) were also retrieved. The full list of features included in the analysis is shown in Table 1.

Table 1: Features included in the model.

| Demographic Characteristics | Age, Height, Weight, Ideal body weight, Body mass index, Sex, Race, Charlson Comorbidity Index, Functional capacity, American Society of Anesthesiologists physical status, Surgery type |
| Comorbid Conditions | Hypertension, Coronary artery disease, Prior myocardial infarction, Congestive heart failure, Diastolic function, Left ventricular ejection fraction, Aortic stenosis, Atrial fibrillation, Pacemaker, Prior stroke, Peripheral artery disease, Deep venous thrombosis, Pulmonary embolism, Diabetes mellitus, Outpatient insulin use, Chronic kidney disease, Ongoing dialysis, Pulmonary hypertension, Chronic obstructive pulmonary disease, Asthma, Obstructive sleep apnea, Cirrhosis, Cancer, Gastro-esophageal reflux, Anemia, Coombs positive, Dementia, Ever-smoker |
| Preop Vital Signs | Systolic blood pressure, Diastolic blood pressure, Pulse oximeter, Heart rate |
| Preop Labs | Albumin, Alanine phosphatase, Creatinine, Glucose, Hematocrit, Partial thromboplastin time, Potassium, Sodium, Urea Nitrogen, White blood cells |
| Intraoperative Time Series | Mean arterial pressure, Systolic blood pressure, Diastolic blood pressure, Heart rate, Pulse oximeter, Temperature, Respiratory rate, Tidal volume, Peak inspiratory pressure, Positive end-expiratory pressure, Fraction inspired oxygen, End-tidal carbon dioxide, End-tidal anesthetic concentration |
| Intraoperative Meds and Fluids | Albumin, Amiodarone, Crystalloid (lactated ringers + normal saline), Dobutamine, Ephedrine, Epinephrine, Fentanyl, Furosemide, Gentamicin, Hydromorphine, Midazolam, Nicardipine, Norepinephrine, Packed red blood cells, Phenylephrine, Propofol, Remifentanil, Vancomycin, Vasopressin, Other blood products |

For intraoperative time series features, summary measures were derived. For each feature, the mean, standard deviation, maximum, and minimum over the entire surgery were calculated. The maximum pulse oximeter reading was omitted due to ceiling effects, while minimum peak inspiratory pressure and minimum tidal volume were omitted due to expected lack of clinical significance. In addition, the fraction of surgery with extreme values of certain parameters were also calculated, using multiple cutoff values. These included duration of low mean arterial pressure (<55, <60, or <65 mmHg), high heart rate (>100, >110, or >120 beats per min), low heart rate (<60, <55, or <50 beats per min), low temperature (<36 or <35.5 °C), low pulse oximeter (<90 or <85%), high exhaled carbon dioxide (>50 mmHg), low exhaled carbon dioxide (<30 mmHg), high peak inspiratory pressure (>30 mmHg), and high tidal volume (>10 mL per kg). Lung compliance was also calculated as final tidal volume divided by final peak inspiratory pressure.

Experimental Technique

For each of the two target outcomes, F-GAM was compared to four baseline models (decision tree [DT], random forest [RF], support vector machine [SVM], and deep neural network [DNN]) and to three GAMs (logistic regression [LR], gradient boosting decision stumps [GBDS] and deep embedding logistic regression [DELR]). Note that density based logistic regression (DLR) was not included as it did not finish training in 24 hours. Each model was trained using a 70% random sample of the dataset. 10% of the dataset was selected as a validation set for hyper-parameter tuning and performance was tested on the remaining 20% of the dataset. Model performance was quantified using area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). We calculate two-sided 95% confidence intervals for each measure using the statistical analysis method given by Hanley and McNeil. 

347
Table 2: AUROC score, AUPRC score and their corresponding 95% confidence interval (CI) of different methods. DT = decision tree. RF = random forest. SVM = support vector machine, DNN = deep neural network. LR = logistic regression, GBDS = gradient boosting decision stumps, DELR = deep embedding logistic regression, F-GAM = factored generalized additive model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acute Kidney Injury</th>
<th>Acute Respiratory Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUROC 95% CI</td>
<td>AUPRC 95% CI</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>0.580 [0.563, 0.597]</td>
<td>0.137 [0.130, 0.145]</td>
</tr>
<tr>
<td>RF</td>
<td>0.820 [0.806, 0.835]</td>
<td>0.253 [0.243, 0.266]</td>
</tr>
<tr>
<td>SVM</td>
<td>0.794 [0.779, 0.809]</td>
<td>0.215 [0.205, 0.226]</td>
</tr>
<tr>
<td>DNN</td>
<td>0.787 [0.772, 0.802]</td>
<td>0.216 [0.206, 0.227]</td>
</tr>
<tr>
<td>GAMs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>0.794 [0.783, 0.813]</td>
<td>0.221 [0.212, 0.233]</td>
</tr>
<tr>
<td>GBDS</td>
<td>0.803 [0.788, 0.818]</td>
<td>0.253 [0.242, 0.265]</td>
</tr>
<tr>
<td>DELR</td>
<td>0.800 [0.786, 0.815]</td>
<td>0.235 [0.225, 0.247]</td>
</tr>
<tr>
<td>Our Method</td>
<td>F-GAM</td>
<td>0.824 [0.813, 0.842]</td>
</tr>
</tbody>
</table>

Results

The dataset included 111,890 patients. Of these patients, 5,018 were excluded from the acute kidney injury model because they were receiving dialysis before surgery or because no postoperative creatinine value was available. Of the remaining 106,872 patients, 6,472 (6.1%) experienced acute kidney injury. Of the original 111,890 patients, 89,688 were excluded from the acute respiratory failure model because they were not admitted to the intensive care unit, while 6,578 were excluded due to preoperative mechanical ventilation or one of the other exclusion criteria. Of the remaining 15,624 patients, 489 (3.1%) experienced acute respiratory failure.

Performance of the models is shown in Table 2, while the receiver-operating characteristic and precision-recall curves are shown in Figure 2. For both outcomes, F-GAM provided the highest AUROC and the highest AUPRC. DT and RF are excluded from Figure 2 for readability purposes.

Figure 3 demonstrates how the contribution $w_t(x^S)f_t(x_t^{TV})$ to the predicted risk of acute kidney injury changes at different values $x_t^{TV}$ of four representative time-varying features in two randomly selected patients. Each panel assumes that all other time-varying features remain constant. Points that are higher on the vertical axis represent a larger contribution to the predicted probability.

Discussion

Our experimental results demonstrate that F-GAM outperforms other methods with respect to accuracy on this task while also offering the benefits of accountability and actionability. All of the models tended to perform better for kidney injury than for respiratory failure, which is likely related to the higher incidence of kidney injury in our dataset and the larger sample size used for this outcome (106,872 versus 15,624). The pure deep neural network didn’t perform as well in our dataset, likely because it is very easy to overfit despite traditional regularization methods such as learning rate decay and weight decay being applied. The random forest model had performance characteristics that were most similar to F-GAM, but F-GAM would be preferable over the random forest because F-GAM offers interpretability, while the random forest does not.
Figure 2: ROC curve and precision recall curve (PRC) of different models predicting acute kidney injury and acute respiratory failure. LR = logistic regression, SVM = support vector machine, GBDS = gradient boosting decision stumps, DNN = deep neural network, DELR = deep embedding logistic regression, F-GAM = factored generalized additive model.

Figure 3 demonstrates how predictions generated by F-GAM can be used to guide intraoperative management. For example, the lower-left panel shows that an increase in maximum heart rate from 100 bpm to 110 bpm appears to be associated with an increased risk for acute kidney injury, because the graph has a steep positive slope in this region. On the other hand, an increase from 80 bpm to 90 bpm is not associated with increased risk, because the graph has a flat slope in this region. Multiplication times the scalar $w(x^3)$ allows the curve to expand or shrink vertically depending on the baseline characteristics and other health conditions of an individual patient. Thus an increase in maximum heart rate from 100 bpm to 110 bpm appears to be associated with increased risk both in a healthy patient (blue curve) and in a very ill patient (orange curve), but the increase in risk (i.e., the slope) is much greater for the very ill patient. This observation fits the anesthesia clinician’s intuition.

The anesthesia clinician should not assume that the associations reported by this model indicate that elevated heart rate causes acute kidney injury. Nor should the clinician assume that blindly giving a medication that lowers the heart rate will decrease the patient’s risk for acute kidney injury. On the contrary, increased heart rate is often a sign of an underlying problem, such as dehydration. The underlying problem, not the fast heart rate, is what increases the risk for acute kidney injury. It is the clinician’s job to identify the underlying problem and correct it.

The upper-left panel provides another example of a scenario where the reported correlation should not be assumed.
Figure 3: Contribution of each feature to the predicted probability of acute kidney injury as a function of feature value. Each panel assumes that all other dynamic features are held constant. The blue curve shows the feature contributions in a 57-year-old healthy female (who ultimately did not have AKI), while the orange curve shows the feature contributions in a 49-year-old female with hypertension, chronic kidney disease, and cirrhosis of the liver (who did have AKI).

to indicate causation. The graph shows a negative slope as the cumulative dose of phenylephrine (a medication that raises blood pressure) increases from 0 to 100 mcg/kg. This suggests that administration of low doses of phenylephrine might decrease the risk of acute kidney injury (if all other features remain constant). This example demonstrates one of the limitations of any factorized model structure; it is unlikely that large doses of phenylephrine decrease the risk of acute kidney injury in and of itself, but it is well supported in the anesthesia literature that untreated low blood pressure increases acute kidney injury risk. Additionally, zero or very low doses of phenylephrine (a weaker first line drug) may represent immediate escalation to stronger vasopressor such as norephinephrine, which increases risk.

In regions where the curve is relatively flat, observing a different value for that feature will have minimal impact on the predicted probability of the target. This can be seen in the left portion of the upper-right panel. Static features (such as age) may also be important (and our model accounts for these effects through the presence of static features in the weights $w_t(x^S)$ and in the bias term), but these features by definition are non-modifiable and change neither when intraoperative problems occur nor when the problems are corrected.

When F-GAM predicts a high probability of an adverse outcome such as acute kidney injury, the time-varying features contributing the most to that prediction are those with the highest current values of $w_t(x^S)f_t(x^TV)$. Curves similar to those shown in Figure 3 can be shown to the clinicians in real time during surgery. When a clinician enters the room to provide assistance with a patient who triggered a high-risk alert, these curves can help the clinician quickly determine what features are important in this particular case. This saves time that would otherwise be spent reviewing all the vital signs and other data. In an environment with as much real-time data as an operating room, streamlining data review is a major advantage, particularly if the clinician coming to provide assistance is otherwise unfamiliar with the patient. Ideally, the clinician will identify the underlying diagnosis sooner and deliver treatment sooner, if treatment is
Conclusion

In this paper, we have described a novel Factored Generalized Additive Model (F-GAM) and demonstrated its use in predicting postoperative acute kidney injury and acute respiratory failure in a historical cohort of patients receiving surgery with anesthesia. F-GAM allows for interactions between static and time-varying input features while retaining the qualities of accountability and actionability. Our model outperformed baseline models and other GAMs in predicting both of the complications tested, and the graphical displays of risk indicated associations that have face validity to an anesthesia clinician. Next steps include application of this technique to other outcomes and prospective deployment of these models for prediction of complications.

References


Implementation of Real-Time Electronic Clinical Decision Support for Emergency Department Patients with Pneumonia Across a Healthcare System

Nathan C. Dean MD, Caroline G. Vines MD, Jenna Rubin RN, Dave S. Collingridge PhD, Mark Mankivsky, Raj Srivastava MD MPH, Barbara E. Jones MD, Kathryn G. Kuttler PhD, Missy Walker MD, Nathan Jenson DO, Brandon J. Webb MD, Todd L. Allen MD, Peter J. Haug MD

Abstract:
A real-time electronic CDS for pneumonia (ePNa) identifies possible pneumonia patients, measures severity and antimicrobial resistance risk, and then recommends disposition, antibiotics, and microbiology studies. Use is voluntary, and clinicians may modify treatment recommendations. ePNa was associated with lower mortality in emergency department (ED) patients versus usual care (Annals EM 66:511). We adapted ePNa for the Cerner EHR, and implemented it across Intermountain Healthcare EDs (Utah, USA) throughout 2018. We introduced ePNa through didactic, interactive presentations to ED clinicians; follow-up visits identified barriers and facilitators to use. Email reminded clinicians and answered questions. Hospital admitting clinicians encouraged ePNa use to smooth care transitions. Audit-and-feedback measured utilization, showing variations from best practice when ePNa and associated electronic order sets were not used. Use was initially low, but gradually increased especially at larger hospitals. A user-friendly interface, frequent reminders, audit-and-feedback, a user survey, a nurse educator, and local physician champions are additive towards implementation success.

Introduction/Background
Pneumonia is an acute infection of the lungs that fills alveoli with inflammatory cells and fluid, making it difficult to breathe, take in oxygen, and expel carbon dioxide. Pneumonia is the eighth leading cause of death in the United States, with more than six million cases annually, one million hospitalizations, and costs over 7 billion dollars for inpatient treatment alone. However, pneumonia care varies dramatically across healthcare providers and institutions. 

Pneumonia treatment is neither straightforward nor simple. When a patient is suspected of having pneumonia, clinicians must (1) assess symptoms and clinical findings and determine whether pneumonia is likely compared to another diagnosis, (2) identify the most appropriate site for treatment, (3) determine whether the bacteria causing the infection is likely to be resistant to commonly prescribed antibiotics, and then select the appropriate antibiotic. Correctly performing these tasks is critical for patient safety but given the complexity of these questions and the fundamental limitations of human decision-making, too often care varies from best practice. For example, without pneumonia clinical decision support (CDS), a provider might discharge a patient home on oral antibiotics instead of arranging for hospital admission to provide needed 24/7 nursing care, electronic monitoring and intravenous antibiotics. Studies have demonstrated variability in rates of hospital admission between different institutions, and two-fold variability in rate of hospitalization between physicians in a single emergency department (ED).

Well-established scientific guidelines exist for pneumonia treatment, but they are underused. Decades of research, quality improvement, and policy work led to the development, deployment, and validation of pneumonia treatment guidelines; most recently, the Infectious Disease Society of America (IDSA) and the American Thoracic Society (ATS) 2007 pneumonia treatment guideline. Guidelines are intended to help clinicians accurately diagnose pneumonia, determine the most appropriate treatment setting, and deliver the most effective treatment. The 2007 guideline was the primary knowledge base behind our real-time, electronic pneumonia CDS tool (ePNa).

Effectively deployed pneumonia CDS should standardize treatment and improve patient safety. Paper-based clinical decision support improves pneumonia treatment, but not enough to fully optimize it. Informed by the 1993 ATS pneumonia treatment guideline, our team deployed a pneumonia treatment CDS tool across Intermountain Healthcare (Utah, USA) starting in 1995 using paper-based flow charts and standardized written order sets. This intervention was associated with reduced mortality for pneumonia patients. Although the results showed effectiveness, the paper-based CDS pneumonia tool did not integrate well into clinical workflow, making widespread adoption difficult. Building on this early work, our team integrated guideline recommendations and best available evidence to develop
ePNa for the Intermountain Healthcare (Intermountain) electronic health record (EHR). ePNa extracts real-time data from the EHR to guide diagnosis, risk stratification, and treatment.

**Favorable Clinical Outcome Data**

ePNa operating in Intermountain’s legacy EHR was deployed in four Salt Lake Valley EDs beginning in May 2011, and during 2012 was used by ED physicians in 63% of pneumonia patients. We reported results of a quasi-experimental implementation study of ePNa in the four EDs compared with three similar Intermountain hospitals along Utah’s urban corridor that served as usual care, concurrent controls. All-cause, 30-day mortality among patients with community-acquired pneumonia where ePNa was available was significantly lower compared to three usual care hospitals (OR: 0.53; 95% CI: 0.28 to 0.99). Unadjusted mortality was 7.8% in usual care hospitals versus 5.2% in the ePNa hospitals. During a baseline year prior to ePNa implementation, mortality of pneumonia patients was similar between the 7 hospitals. Patients with higher severity illness were more likely to be admitted to ePNa hospitals for treatment versus control hospitals where more of those patients were sent home. Patients in the ePNa hospitals were more likely to receive guideline recommended antibiotics.

**Goals**

Based on the value of ePNa demonstrated in the four Salt Lake Valley EDs, the Intermountain Office of Research funded a large-scale effort to develop and deploy ePNa for Intermountain’s version of the newly installed Cerner EHR (iCentra) across all 22 Intermountain hospitals. The deployment of ePNa was designed and conducted in step-wise fashion according to a prespecified implementation trial (ClinicalTrials.gov Identifier: NCT03358342). Our goals were:

1. Adapt ePNa from its legacy version to function well in iCentra. To accomplish this, we wanted to learn from clinician user observations, conduct a formal survey, and incorporate their suggestions to continuously improve ePNa function.
2. Achieve 80% ED physician utilization of ePNa in patients diagnosed with pneumonia at Intermountain hospitals. We targeted 80% rather than 100% use because of diagnostic uncertainty during ED patient care, and to be realistic considering ePNa’s “opt in” design.
3. Identify effective methods of implementing CDS for EDs, including barriers and facilitators.
4. Measure processes of care and clinical outcomes associated with implementation of ePNa after trial completion.

**Description of ePNa**

ePNa is an innovative, real-time electronic CDS tool for pneumonia treatment in the ED. No comparable tools have been described in published literature. It is unique in having been deployed and utilized by ED clinicians with published favorable clinical outcome data.

ePNa logic incorporates the Five Rights Framework. (Table 1) ePNa integrates a pneumonia detection system with a management tool that delivers the right information to ED clinicians caring for patients with suspected pneumonia. Its detection system helps identify ED patients who might have pneumonia based on their presenting symptoms, physical exam, laboratory and radiographic findings. The clinical management system of ePNa evaluates each patient’s severity of illness to recommend (1) which patient disposition (outpatient care, inpatient ward care, or the intensive care unit) is the safest setting for pneumonia treatment, (2) which antibiotics should be prescribed, and (3) which microbiology studies (e.g., blood culture) are warranted. Patients with lower severity diseases treated at home have been shown to be more satisfied with their care and to return to work and/or usual activities more rapidly compared with similar severity patients admitted to the hospital. Direct cost of care is 20 times higher for comparable patients admitted to the hospital versus those treated at home. However, more severely ill patients are more safely treated by direct admission to a hospital ward or intensive care unit.

<table>
<thead>
<tr>
<th>Right</th>
<th>ePNa Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Deliver the right information</td>
<td>Chief complaint, vital signs, chest imaging, age, laboratory findings, microbiology results and antibiotic resistance information</td>
</tr>
<tr>
<td>2. To the right person</td>
<td>Clinicians at the point of pneumonia care</td>
</tr>
<tr>
<td>3. Using the right format</td>
<td>Simple, single recommendations for diagnosis, treatment, and disposition per screen with accompanying supportive data.</td>
</tr>
<tr>
<td>4. In the right channel</td>
<td>Embedded electronically in the health record</td>
</tr>
<tr>
<td>5. At the right time</td>
<td>From the beginning, throughout, and the end of the pneumonia episode of care.</td>
</tr>
</tbody>
</table>

Table 1: ePNa and the Five Rights Framework of Decision Support
effectively communicated. Ideally, initial ED treatment comprises best practice that is smoothly transitioned and continued in the hospital. However, if the inpatient team makes different treatment decisions based on individual clinician variability rather than patient-centered illness features, there is discontinuity of care and potential for patient harm. Similarly, patients returning for follow-up a day or two following their initial ED encounter are often seen by a different clinician. Standardizing clinical practice facilitates continuity of care by different providers.

**Technical description of ePNa**

ePNa consists of a pneumonia detection system combined with a management tool seamlessly integrated into usual ED physician workflow through a single user interface.

**Detection system**
The ePNa detection system is embedded in the iCentra decision support environment. It leverages tools designed to run prediction and classification algorithms in a clinical setting. It operates by consuming a stream of 40 clinical data elements captured from the EHR during ED patient care (Figure 1). The ePNa screening framework uses a natural language processing (NLP) subsystem to identify information in free-text reports to determine the presence of pneumonia. The detection system relies on a probabilistic Bayesian network to calculate likelihood of pneumonia.

When the combination of data results in a probability of pneumonia ≥ 40%, ePNa alerts the clinician using asynchronous messaging. At this point, the clinician can launch the tool, but may opt out if the patient is not believed to have pneumonia. Physicians may identify patients with a calculated risk of < 40% estimated pneumonia likelihood and still activate the management tool. The 40% threshold yields a 50% positive predictive value of the pneumonia alert, thereby decreasing “alert fatigue” and limiting negative impact on the ED clinical environment. After confirmation of pneumonia, ePNa takes the clinician through risk-stratification and recommends patient disposition and treatment. Physicians may also directly launch the management portion of ePNa without an alert, utilized when calculated pneumonia probability is < 40%.

**ePNa objectively calculates severity of illness for each patient with pneumonia.**

1) Estimates 30-day mortality risk using an electronic version of a validated severity score (eCURB). The 5 elements are patient age, initial systolic blood pressure, initial respiratory rate, altered mental status, and blood urea nitrogen level.

2) Calculates the patient’s ability to oxygenate arterial blood compared to the fraction of inspired oxygen (PaO2/FiO2) from measured percutaneous oxygen saturation (SpO2).

3) Gathers electronically the 9 minor criteria for severe pneumonia (IDSA/ATS 2007), summed into an ordinal score.

4) Identifies patients with an associated pleural effusion by NLP of the chest imaging report

ePNa then recommends (1) patient disposition, (2) antibiotic therapy, and (3) suggested microbiology tests appropriate to the patient’s condition. It displays the data in a form suitable for physician review. The physician can alter tool recommendations based on individualized clinical judgment.

**Site of care - outpatient, hospital ward, or intensive care unit**

Patients with eCURB-predicted mortality ≥ 5%, PaO2/FiO2 < 280 mm Hg adjusted for altitude, or significant pleural effusion are recommended for hospital admission in accordance with previously validated criteria. Patients with ≥
3 severe Community Acquired Pneumonia (CAP) minor criteria, or PaO2/FiO2 < 120 mm Hg are recommended for direct admission to the intensive care unit. (figure 2) Severity information is displayed to the ED physician with the site of care recommendation. The open loop design of ePNa allows the physician to accept or reject this recommendation; if rejected, the physician is asked to provide a reason from a pre-populated pick list or to add free text. This functionality allows real-time learning from the use of the tool.20

Identification of patients with risk factors for antibiotic resistant bacterial pathogens
Drug Resistance In Pneumonia (DRIP) is a 10 factor ordinal score that measures risk for antibiotic resistant bacterial pathogens developed at Intermountain and externally validated.21 In 2014, ePNa was updated to electronically calculate DRIP and incorporate its results into antibiotic treatment recommendations. This operation involves the following:

1. Searches the longitudinal EHR for hospitalization ≥48 hours within the previous 60 days and matches facility name and addresses for skilled nursing homes, rehabilitation facilities, and long-term acute care facilities. It also searches International Classification of Diseases codes version 10 (ICD-10) indicative of chronic obstructive pulmonary disease, diabetes, chronic heart failure, and active cancer.
2. Extracts previous microbiology results for Methicillin-resistant Staphylococcus aureus and Pseudomonas aeruginosa.
3. Searches electronic prescribing records for antibiotic use within the prior three months or use of gastric acid reducing medications.
4. Applies NLP to ED chest imaging reports to identify presence of a feeding tube, a measure of functional status.
5. Asks the ED physician to confirm that the patient is ambulatory (versus chair/bed bound).

Figure 3 shows the user interface for determining the DRIP score. Physicians click on the right-sided box to add or delete risk factors if there is information not present in the electronic health record (e.g. a visitor to Utah without
Diagnostic testing recommended by ePNa

Patients admitted to the Intensive Care Unit are recommended for two blood cultures, urine antigen tests for *Streptococcus pneumoniae* and *Legionella pneumoniae*, nasal swab for *Methicillin resistant Staphylococcus aureus* and viral respiratory pathogens by Polymerase Chain Reaction, and a tracheal aspirate for semi-quantitative culture if newly intubated. The same battery of studies is recommended for patients with a DRIP score >3 being admitted to a hospital ward.

Adaptation of ePNa for iCentra

Beginning in 2016, Intermountain changed over from its legacy EHR to iCentra. Beginning in January 2017, ePNa was reprogrammed and adapted to operate within iCentra. Beta testing began in the 4 Salt Lake Valley EDs in July 2017. After implementation in a 5th ED in November 2017 identified additional issues with functionality requiring reprogramming, a successful re-launch began in January 2018 across the remaining Intermountain Hospitals.

The iCentra version of ePNa retained a “P” alert on the ED electronic tracking board (LaunchPoint), but the one-click launch functionality of the legacy version could not be replicated. Three “clicks” from the physician user are needed to launch ePNa in the iCentra version, but then ePNa functions easily once initiated. The iCentra version of ePNa features a direct link to Computerized Physician Order Entry that was not available in the legacy EHR. As the last step of ePNa, the clinician launches one of 6 pneumonia order sets (2 outpatient, 2 hospital ward, 2 intensive care unit, each with a low risk/high risk for antibiotic resistant pathogens version). ePNa iCentra also contains drug allergy information linked to the order entry set, functionality not available in the legacy EHR.

Automatic loading of ePNa output into clinician ED and admission notes. ePNa’s calculated percent likelihood of pneumonia, the severity calculations (eCURB, PaO2/FiO2, and severe CAP criteria), and the DRIP score are automatically loaded into clinician documentation under a sectional heading of “Clinical Decision Support.” In addition, ePNa’s recommendation for site of care, whether the ED physician agreed with its recommendations, plus any additional ED physician comments, are automatically loaded in ED and admission notes.
**Implementation of ePNa across Intermountain**

Intermountain operates hospitals in small, rural locations (often staffed by family practice physicians) as well as large, tertiary care medical centers in urban regions that use a board-certified emergency medicine staffing model. After achieving acceptable tool functionality in January 2018, we began implementing ePNa into 6 geographic clusters of Intermountain Hospital ED’s at approximately 2-month intervals. We began with larger hospitals, with the later implementation sites being smaller rural hospitals.

**Methods of CDS implementation**

1) **Front line provider education:** ED physicians, Advanced Practice Clinicians, ED nurses, admitting hospital and critical care physicians first received didactic, customized and interactive presentations prior to ePNa launch. Emergency departments were the focus for implementation since if ePNa isn't run in the ED, it is less likely to be started after hospital admission. Emergency department 'buy-in' is key and promotes continuity of care between providers throughout the hospital stay.

2) **Nurse educator engagement:** Multiple follow up visits by the ED nurse educator (JR) began after ePNa launch at each site.

3) **Physician champions:** We identified physician champions in every ED to teach and encourage their peers to adopt ePNa into routine care of pneumonia patients. The largest ED physician group identified thirteen members as super users available to answer questions and promote utilization while working in the ED.

4) **Feedback to improve the CDS tool:** Physician champions provided a voice for feedback and insights to functionality of their location. Individual providers also gave feedback to the nurse educator during her visits. Follow up group meetings were conducted with physicians from the ePNa team, answering questions and concerns from providers.

5) **Email reminders:** Reminders and answers to frequently asked questions were sent to providers.

6) **Audit and feedback:** Repeated audits of ePNa use were fed back to providers along with issues involving pneumonia patient care where ePNa had not been used. Audits that included the physician provider and patient encounter identifiers were made available to all providers at each ED, engendering discussion and provider accountability. We observed the largest increases in ePNa utilization after audit results were openly shared by local champions/ED leadership, leading to discussion among providers.

7) **Admitting hospital physicians “encourage” the routine use of ePNa by ED physicians and increase patient safety by facilitating transitions of care from the ED to the hospital.** ED physicians frequently comment that ePNa structures and supports their conversations with admitting physicians.

**Rate of ePNa uptake**

Despite published favorable outcome data, ED physicians were initially slow to adopt the iCentra ePNa, judging its value by the user interface and subjective assessment of its impact on their individual clinical care. This adoption pattern was more pronounced outside of Salt Lake Valley (where clinicians had used the legacy version of ePNa). Audits were conducted over time to measure uptake and use of the tool. These audits showed variations from best practice in about half of pneumonia patients when ePNa was not used, or when the associated computerized order set was not also used. Six audits were performed during 2018 (Figure 4). Hospital clusters where ePNa was launched earlier in 2018 have more measurements than later clusters. Each audit spanned 2 to 4 weeks (longer at smaller hospitals to increase the number of patients), ranging between 4 and 30 cases per facility depending on the interval and facility ED case volume. Patients were initially identified by ICD-10 codes for pneumonia, or for sepsis and acute respiratory failure with pneumonia present as a secondary diagnosis. Physicians (ND, CV, NJ, MW) individually reviewed each case to verify ED physician diagnosis of pneumonia, excluding those where the physician diagnosis was not pneumonia, e.g. aspiration pneumonitis. Figure 4 illustrates the percent utilization of ePNa by hospital over the 6 audits.
Results of emergency department physician survey.
A REDCAP confidential survey of ED physicians was conducted in December 2018. Survey details have been previously published. The response rate of completed surveys was 53%, or 90 of 169 invited physicians. We used Principal Component Analysis followed by Reliability testing to group like questions together. Figure 5 illustrates respondents’ opinions about how well ePNa helps with clinical decision-making. The mean score on component 1 is 3.6 (SD=0.79). Response options ranged from strongly disagree (1) to strongly agree (5) with a neutral rating corresponding to (3), a mean of 3.6 indicates that the majority of respondents hold favorable opinions about how well ePNa helps their clinical decision-making. However, responses regarding decision support with diagnosing pneumonia revealed that most respondents felt ePNa does not help. Respondents’ opinions of the “P” alert system (component 2) resulted in a mean score of 3.1 (SD=1.3). A mean of 3.1 is approximately in the center of the scale. However, scores are spread out across the response scale, indicating that respondents...
have diverse opinions on how well the alert system works.

The median response to questions about how much time ePNa requires is 3.0 (IQR=2) (Figure 6). This median is in the center of the scale, suggesting that respondents are split over how much time it takes to use ePNa.

The median response for “ePNa is free from technical problems” is 2.0 (IQR=1.0), Figure 7. A median of 2.0 is on the disagree side of the scale, suggesting that most respondents feel that ePNa does have technical problems.

With the mean responses to components 1 and 2 being below 4.0 (agree), progress can be made in improving respondents’ perceptions toward ePNa. The comments on improving ePNa support this conclusion. Removing bugs and glitches and adding new functionality will likely enhance perceptions of ePNa.

There were a few strong, negative comments about ePNa. However, these comments were balanced by an equal number of strong, positive comments. People with strong, negative opinions may never be persuaded that ePNa is effective and usable no matter how well it works, or that CDS is needed. However, in this survey the majority of respondents believe that ePNa is slightly above average in quality, and those opinions may improve with enhancements to the interface and functionality of ePNa. Efforts on improving perception and utilization should focus on clinicians who are amenable to adopting CDS into their clinical care.

Two survey comments from clinicians represent several similar comments and illustrate current limitations of ePNa.

“The tool often gives recommendations that are inappropriate and potentially dangerous. I’ve had it recommend outpatient therapy for patients who need the ICU and vice versa.” ePNa requires accurate, real-time data to calculate severity of illness which drive disposition recommendations. The nurse educator has emphasized to ED nurses the importance of charting SpO2 on room air (for calculating PaO2/FiO2), and confusion (oriented to person, place, and time or not; key components of both eCURB and the severe CAP criteria). For example, a patient brought in by ambulance may have had a room air SpO2 of 60% measured by prehospital medics and placed on high flow oxygen by mask, resulting in a saturation of 99% on arrival to the ED. Only the latter might be available to ePNa, but PaO2/FiO2 is not accurately calculated from high SpO2.
Uncharted data are considered normal by all severity systems, yet if confusion is present and not charted, ePNa might give a “dangerous” recommendation. ePNa requires timely, accurate entry of patient data, yet the frequently busy and occasionally chaotic ED environment can make accurate CDS function difficult. ePNa is open loop CDS which cannot anticipate every clinical scenario, thus the treating physician must do what he/she feels is safe and appropriate for the individual pneumonia patient.

Another illustrative comment: “Better alert system. A dedicated care pathway column would make the user interface and inter-operability much improved”. The screening algorithm was not updated for iCentra ePNa, and due to technical limitations, the single click for launching ePNa featured in the legacy version could not be carried over. This presented an early barrier to use and adoption of ePNa. We are now re-deriving the screening tool for iCentra and plan a pop-up dialog box to launch the program with a single click once ePNa identifies a patient with ≥ 40% likelihood of pneumonia.

Summary and Conclusions

ePNa is an innovative electronic CDS tool for ED patients with pneumonia. Published outcome data demonstrated reduced mortality and appropriate hospital admission of high severity patients with pneumonia. A long-term goal of the current study is to demonstrate similar results across the large and small, urban and rural Intermountain Hospitals. However, the huge challenge of broad adoption by ED physicians must occur first.

Adaptation of ePNa into iCentra allowed implementation across all Intermountain hospitals. In this paper, we have described our implementation process, reported ePNa usage during 2018, and results of our ED physician survey. Physician utilization has been higher in larger, urban medical centers with longer exposure to ePNa. Recurring personal and electronic feedback/reminders from the ePNa team is increasing its use across Intermountain. Improvements in the pneumonia detection component and improving the user interface that launches ePNa are underway, as identified by ED physician users and described above. More work to improve charting of critical data elements by ED nurses will increase tool accuracy and use of ePNa. We hope to eventually export ePNa outside of Intermountain to determine its generalizability among different pneumonia populations and providers with different diagnosis and treatment patterns. These efforts will benefit from the lessons learned during the implementation of ePNa at Intermountain.

References

11 Campbell R. The five “rights” of clinical decision support J AHIMA 2013 84(10):42-7
Development of a Genomic Data Flow Framework: Results of a Survey Administered to NIH-NHGRI IGNITE and eMERGE Consortia Participants

Paul Dexter, MD,1,2 Henry Ong, PhD,3 Amanda Elsey, MHA,4 Gillian Bell, PharmD,5 Nephi Walton, MD, MS,6 Wendy Chung, MD, PhD,7 Luke Rasmussen, MS,8 Kevin Hicks, PharmD, PhD,9 Aniwaa Owusu-obeng, PharmD,10 Stuart Scott, PhD, FACMG,10 Steve Ellis,10 Josh Peterson, MD, MPH3

1Indiana University School of Medicine, Indianapolis, IN; 2Regenstrief Institute, Inc., Indianapolis, IN; 3Vanderbilt University Medical Center, Nashville, TN; 4University of Florida, Gainesville, FL; 5Mission Health, Asheville, NC; 6Geisinger Health System, Danville, PA; 7Columbia University, New York, NY; 8Northwestern University, Chicago, IL; 9Moffitt Cancer Center, Tampa, FL; 10Icahn School of Medicine at Mount Sinai, New York, NY

Abstract

Precision health's more individualized molecular approach will enrich our understanding of disease etiology and patient outcomes. Universal implementation of precision health will not be feasible, however, until there is much greater automation of processes related to genomic data transmission, transformation, and interpretation. In this paper, we describe a framework for genomic data flow developed by the Clinical Informatics Work Group of the NIH National Human Genome Research Institute (NHGRI) IGNITE Network consortium. We subsequently report the results of a genomic data flow survey administered to sites funded by NIH-NHGRI for large scale genomic medicine implementations. Finally, we discuss insights and challenges identified through these survey results as they relate to both the current and a desirable future state of genomic data flow.

Introduction

Precision health's more individualized molecular approach will enrich our understanding of disease etiology and patient outcomes.1 Individuals receiving treatment within healthcare systems will ultimately be characterized by multiple methods, including genomics, epigenomics, metabolomics, and proteomics.1

Acquiring and interpreting these new data sources are now possible because of the decreasing costs of molecular testing, development of large-scale biologic databases, and computational tools that facilitate analysis of large data sets.1 Genetic sequencing in particular is on the verge of becoming a routine part of clinical care.2 However, major challenges to routine use persist. Variant interpretation and reporting is currently extraordinarily complex and time-intensive.3 Genomic lab instrument interoperability with clinical information systems is hindered by a lack of adopted health IT standards.3 Finally, clinical decision support (CDS) necessary for clinicians to understand the significance of all clinically important variants is not widely implemented. At every step, from the point of raw data being generated by the genomics instrument to CDS at the point of care, the delivery of genomic data will require markedly improved automation.

This manuscript focuses on the current status of genomic data transmission, transformation, and interpretation (flow) from instrument to point of care. Specifically, we describe a framework for genomic data flow developed by the Clinical Informatics Work Group (CIWG) of NIH-NHGRI's "Implementing genomics in practice" (IGNITE) Network. We subsequently report the results of a related survey administered to sites funded by NIH-NHGRI for large scale genomic medicine implementations. Finally, we discuss insights and challenges identified through these survey results as they relate to both the current and a desirable future state of genomic data flow.

Methods
Over the course of several months and through a series of meetings, members of IGNITE's CIWG committee developed a draft framework for genomic data flow informed by their own institutional experiences and the medical literature. This draft framework was presented to the full IGNITE consortia and refined based on further feedback.

Members of the CIWG Committee subsequently developed a RedCap survey that corresponded closely to components of the draft framework. For purposes of our survey, we focused on germline molecular testing, but we anticipate that many aspects would be generalizable to somatic molecular testing. Given our primary goal of refining a genomic data flow framework, we chose to survey institutions with well-developed advanced genomic molecular testing laboratories. To this end, we invited survey respondents from institutions that had been funded by NIH-NHGRI as part of the IGNITE and Electronic Medical Records and Genomics (eMERGE) consortia to implement large scale genomic medicine implementations. Surveys were conducted from May 2018 to February 2019.

In the following discussion, we illustrate aspects of the data architecture using two important examples of genomic molecular testing, whole genome sequencing (WGS) and pharmacogenetics (PGx).

Results

Development of a genomic data flow framework. The final proposed framework for genomic data flow is depicted in figure 1. For the sake of simplicity, in the descriptions of the components of figure 1, we will refer only to the number (e.g., "component #2") rather than a full description (e.g., "refer to component #2 of figure 1").

In figure 1, genomic data generally flows left to right. Genomic instrument results from the institution's own laboratory are processed and transformed by a bio-informatics pipeline (component #4). All identified variant results for a patient (component #7) are stored for at least the life of that patient, and separately, interpretative annotations for those results are stored (component #8). For purposes of electronic health record (EHR) performance, clinically actionable subsets of variant data (component #10) and the corresponding interpretative annotations (component #11) are stored, which are available through the EHR (component #13) and for CDS (component #12).

For pathology reporting and reasons discussed below, we distinguish between automatically-generated portions of a draft report (component #5) and the final report completed by the laboratory's genetics expert (component #6). A genome variant knowledge base (component #3) is required in machine-readable form for purposes of the bio-informatics pipeline and automatically-generated portions of a draft report, as well as in human-readable form for the genetics expert. Discrete results and annotations data from an external genomics laboratory (component #1) flow into the full variant (component #7) and interpretative annotations (component #8) databases, while the external pathology report would flow through the pathology reporting system (component #6) and into the EHR (component #13). Finally, processes would allow for ongoing re-interpretation of results as genomic knowledge evolves (component #9) and genomics results would be shared with the patient through a patient portal (component #14). These framework components are discussed more fully below.
Genomic laboratory instrument (internal and external testing). Genomics laboratory instruments rely on a variety of underlying laboratory techniques (e.g., DNA arrays, DNA sequencing, real-time quantitative polymerase-chain-reaction). Such testing is performed in a variety of clinical contexts e.g., prenatal testing or determining predisposition to cancer). For purposes of this paper, our focus begins with the raw data generated by the instruments for genomic tests performed both internally within the institution's laboratory (component #2) and at an external laboratory (component #1).

In the case of WGS, raw data typically consist of fluorescent, chemiluminescent, or electrical current signals. These are transformed into sequential base calls using platform-specific algorithms. The resultant FASTQ file can contain the results for millions of short DNA sequences. It is this FASTQ file that provides the source input for platform-independent software processing, as discussed below in the bio-informatics pipeline section.

These sequence reads are appropriately associated with metadata that describe the methods and technologies used to produce those results. In the case of WGS, such metadata commonly include sequencing quality scores. These quality scores are important for interpretation given that they reflect the statistical confidence that a given base call is correct.

Many genomics laboratory instruments were not designed to work in a clinical networked environment. There is need for both greater interoperability and implementation of standards. To such ends, the HL7 Clinical Genomics Work Group recommends a transcoding process, whereby genomic data are transformed from bio-informatics format into healthcare IT data standards. They also describe an alternative approach where genomic data are encapsulated in healthcare standards.

Bio-informatics pipeline. As described in Leipzig et al.’s review of bio-informatic pipeline frameworks, a bio-informatics pipeline invariably involves the shepherding of files through a series of transformations. These pipelines have evolved into frameworks that accommodate integration of user-defined tools, definition of both serial and parallel steps, complex dependencies, and varied data file types.

In the case of WGS, the source input for a platform-independent bio-informatics pipeline is commonly in the form of a FASTQ file, which can contain millions of short DNA sequences. Through the pipeline, these short sequences are aligned to a human genome reference sequence. Through comparison to the reference sequence, variants for an individual are also determined ("called"). Variants can include single nucleotide variants, insertions/deletions, and copy number variants. Identified variants are prioritized, and then annotated with respect to the clinical relevance of those particular variants. Throughout the process, the data are transformed into successively smaller file formats, such as FASTQ, BAM, SAM, and finally VCF. Pipeline processing for WGS is frequently time-intensive. Both open source and proprietary software is available for these transformations (e.g., GATK, SAMtools, Atlas2).

From an abstract level, pipeline processing for PGx genotyping is similar, insofar as variants are called, prioritized, and annotated for clinical relevance. PGx pipelines commonly rely on translation tables. Such tables relate diplotype results (e.g., CYP2C19 *2/*2) to drug related phenotypes (e.g., poor metabolizer) and CDS recommendations (e.g., "Increased risk for reduced response to clopidogrel. Consider alternative drug").

Such prioritization and curation of variants have become immensely more challenging as a result of WGS and its large number of associated variants. For example, the typical human genome has more than 4 million variants. As noted by Kohane et al., the burden of false-positive incidental findings in WGS "threatens current capabilities to deliver clinical-grade whole-genome clinical interpretation."

Genomics pathology report with interpretation (automatically generated and final). The quality of final genomics pathology reports is a key determinant of the effectiveness of genetic medicine, with high-quality reports concisely explaining both the variants identified in a patient and their clinical relevance. Producing these reports is also labor-intensive and expensive.

While the actual sequencing is getting more affordable and higher quality, efficient interpretation is lagging. In one study of the clinical implications of WGS, Dewey et. al found that curation of the 90 to 127 genetic variants found in
each participant required a median of 54 minutes per genetic variant\textsuperscript{17} (i.e., presumably approximately 100 hours of interpretation by genetic experts per sequenced participant).

Given our belief that this level of intensity for manual curation for WGS would not accommodate universal WGS testing, we distinguish between the automatically-generated portions of a draft pathology report (component #5) and the final pathology report completed by the laboratory's genetics expert (#6). Over time, and out of necessity, we anticipate that the automatically-generated portions of the report will become more complete and valid. Through interfacing to the EHR, it also seems likely that such automatically-generated reports will incorporate an increasing amount of patient-specific clinical context (including family history). While it may never prove completely possible due to inherent complexity, the goal would arguably be for valid WGS results to require no more manual expert curation than is currently associated with complete blood count (CBC) results.

\textit{Genomic variant knowledge base.} A genome variant knowledge base (component #3) is required in machine-readable or computable form for purposes of the bio-informatics pipeline and automatically-generated portions of a draft pathology report. A human-readable form is required for the genetics expert who is completing the pathology report. Such a genomic variant knowledge base could also be employed for purposes of CDS logic (e.g., if variant patterns for particular conditions are coupled to actionable recommendations).\textsuperscript{18}

Common variant knowledge bases include the Database of Single Nucleotide Polymorphisms (dbSNP) Online Mendelian Inheritance in Man (OMIM), and ClinVar.

\textit{Full variant database with meta-data and a separate interpretative annotation database.} Given that germline genomic results are applicable for the life of an individual, and the recognized clinical significance of those results will change as a result of evolving knowledge, a full genome database consisting of all variants and metadata requires long term storage.\textsuperscript{18} Long term storage of a full variant database including metadata (component #7), and a separate interpretative annotation database (component #8) is consistent with at least three of Masys et al.'s desiderata: (1) maintain separation of primary molecular observations from the clinical interpretations of those data, (2) maintain linkage of molecular observations to the laboratory methods used to generate them, and (3) anticipate fundamental changes in the understanding of human molecular variation.\textsuperscript{5}

Given nascent EHR support for genomic data, and the large size of many of these full variant files (e.g., a FASTQ file can be 250 gigabytes for WGS\textsuperscript{12}), these databases typically exist outside of an institution's EHR. Rapid retrieval and analysis of such full genomic data is challenging for the typical EHR database system and could be expected to lead to performance issues.\textsuperscript{19} While such files can be compressed, Masys et al. have emphasized the importance of lossless data compression with the ability to produce a fully accurate copy of the original sequence.\textsuperscript{5}

\textit{Clinically relevant subset of variant database and separate interpretative annotation database.} Similar to the case of the full variant database and their interpretations, and consistent with Masys et al.'s desiderata,\textsuperscript{5} our proposed framework separates the subset of clinically relevant variants (component #10) from their interpretative annotations (component #11).

Recommendations for storage of a clinically actionable subset of variants are based on the fact that "the amount of molecular sequence data that currently has demonstrated clinical significance is a tiny fraction of the full genome and proteome, and it is neither computationally feasible nor desirable to query or analyze one's entire genome in real time to support healthcare-related decisions."\textsuperscript{5} Such a clinically actionable subset would be accessible to the EHR and CDS.

Such clinically relevant subsets of variant information seem to be commonly stored in the EHR problem list as a form of work-around for ensuring that clinicians are aware of the clinically actionable variant information for a patient, and for driving CDS.\textsuperscript{13,20} As EHR vendors improve their systems' ability to incorporate discrete genomic data, similar to how patient drug allergy information is not typically embedded in problem lists, it is likely that variant data will be both displayed to clinicians and available to CDS in non-problem list sections of the EHR.
Ongoing re-interpretation of individuals’ genomic results as knowledge evolves. Germline genetic results have lifelong ramifications for individuals. While an individual’s germline genomics is expected to remain immutable, the clinical significance of that individual’s identified variants will change dramatically over the foreseeable future as a result of ongoing intense genomics research. Need for corresponding ongoing re-interpretation of genetic results (component #9) makes a compelling case for standards-based, structured (computable) electronic reports.8 Both the pathology reports and discrete data that drive CDS will require updating.

As described by Aronson et al., the GenelInsight Suite was specifically designed to accommodate this need for ongoing re-interpretation.16 It relies on a centralized genetic knowledgebase and a highly flexible report-generation tool. A single action by a geneticist updates both future reports and generates alerts for clinicians treating patients. Automatically-generated draft reports and user-defined templates save geneticist time. Clinical alerts are driven off of changes to the genetic knowledge base.

EHR with integrated CDS. As of 2013, there were over 2500 clinical genetic tests available to clinicians.21 Even more overwhelming, the average human genome contains over 4 million variants. The full vision of precision health and genomics (and metabolomics, proteomics, etc) will not be realized in the absence of robust computerized CDS. Until such time as every variant for every person is stored and accessible, precision health will require both pre-test CDS (triggered when a clinician takes an action that should be informed by a genetic assessment but there is no record of such) and post-test CDS (alerts triggered when an action is taken that may be contraindicated by a patient’s genetic profile).22

Given current health care infrastructure, the most straightforward method for robust genomics CDS at the point of care on a widespread basis would currently be to leverage existing EHRs (component #13) with integrated CDS (#12) – potentially in the form of CDS web services. Welch et. al have similarly argued that a CDS architecture would appropriately primarily rely on EHR capabilities that are either currently supported or likely to be supported in the near future.18

Patient portal. The patient is a central stakeholder with respect to precision health. It has been hypothesized that modules incorporated into the patient portal may substitute for or lessen the burden of genetic counseling.23 Patient portal applications and sharing of results may also increase patients’ participation in their own care.23 For such purposes, both discrete and final report data would be appropriately shared with the patient through a secure patient portal (component #14).

Survey results

Genomic laboratory instrument (internal testing). 86% (6/7) of survey respondents reported that their site performs CLIA-certified germline molecular testing internally. Table 1 lists the types of CLIA-certified germline molecular testing internally performed at these six sites. Table 2 reports the clinical contexts for which internal germline molecular testing occurs at these 6 sites. In all cases, reports related to internal molecular testing are available for clinician review within the EHR.

60% (3/5) of sites reported that these reports are stored as PDFs in the EHR, while 20% (1/5) reported storage in the structured XML-based Clinical Document Architecture (CDA) format and 20% (1/5) reported storage in a proprietary format. In the case of pharmacogenomics testing, 75% (3/4) of sites reported that discrete genetic results are transmitted from the laboratory instrument to the EHR.

Discrete genetic results are transmitted much less frequently from the instrument to the EHR in other clinical contexts: 0% (0/1) of testing for genetic cancer predisposition, 0% (0/5) of perinatal genetic disorder testing, 25% (1/4) of bone marrow and organ transplant testing, and 0% (0/3) metabolic and vascular disease marker testing.
Genomic laboratory instrument (external testing). 100% (7/7) of survey respondents reported that their site sends at least a portion of their germline molecular tests to an external CLIA-certified laboratory. Table 2 reports the clinical contexts for which such external germline molecular testing occurs at these 7 sites. In all cases, these external reports are available for clinician review in the EHR, almost exclusively either as a PDF or scanned image. The exception was one site that stored their externally-provided pharmacogenomics report in structured CDA format.

For most types of external testing, the results of external germline molecular testing are stored only as a report in the EHR. In the case of external pharmacogenomics testing, discrete genetic data are stored as a result of external pharmacogenomics testing at 40% (2/5) of sites. In the case of external determination of genetic predisposition to cancer, discrete data are stored at 17% (1/6) of sites. No discrete genetic data are stored as a result of genetic disorders tested perinatally, bone marrow and organ transplants, to detect metabolic and vascular disease markers, or testing for neurodevelopmental disorders.

Bio-informatics pipeline. 86% (6/7) of sites support a computer-based genomics data pipeline platform. Among these six sites, pipelines are employed for genotyping (83%) and sequencing (67%). The most common reason for supporting a computer-based pipeline among these 6 sites is for purposes of pharmacogenomics (83% of sites). Internal personnel maintain these pipelines at all six sites. 50% (3/6) employ an open source pipeline framework, 33% (2/6) employ a commercial framework, and 67% (4/6) employ custom scripting. 50% (3/6) utilize PGx translation tables.

Respondents noted challenges related to these pipelines, including (1) the need to upload the pharmacogenomics phenotype information into the EHR as a discrete field, (2) updating and maintaining the underlying data related to newer medical literature, and (3) non-conformity of different methods of variant calling.

For the six sites that perform internal germline molecular testing, structured variant data is stored outside of the genomics instrument at 67% (4/6) of sites. In all such cases, these structured variant data are stored external to the EHR. For 75% (3/4) of sites, annotations are stored with the structured variant information. For 50% (2/4) of sites, a subset of the variants are available to the EHR (e.g., through entry into the patient's problem list). There are no sites where all variants are accessible within the EHR.

Genomics pathology report with interpretation (automatically generated and final). 100% (6/6) of sites that perform internal germline molecular testing create interpretative reports. 83% (5/6) of sites automatically generate a portion of the results (e.g., variant calls) on this report. 100% (6/6) automatically generate a portion of the interpretation (e.g., phenotype for PGx variants) on this report. Genetics experts utilize templates to create these reports at 100% (6/6) of sites.

Genomic variant knowledge base. Genetic experts at these six sites rely on Pubmed (100% of sites), PharmGKB (83% of sites), dbSNP (83% of sites), CPIC guidelines (83% of sites), laboratory-maintained resources (67% of sites), OMIM (67% of sites), RefSeq (67% of sites), GeneTests (33% of sites), and other resources (33% of sites). 67% (4/6) of sites noted time-consuming aspects associated with interpretation, including (1) NGS with multiple variants can "easily take 4-8 hours of manual effort," (2) integration of pharmacogenetic test results with other clinical factors, and (3) rare variants and variants with incomplete penetrance require extensive curation.

Ongoing re-interpretation of individuals' genomic results as knowledge evolves. 86% (6 of 7) of institutions reported updating prior interpretations as a result of new variant knowledge. At these 6 sites, such updates have been "triggered" by clinician request (50% of sites), patient request (17% of sites), and as part of a scheduled review process (33% of sites). Updates have originated from the internal laboratory (100% of sites), as well as the external laboratory (50% of sites).
Efforts to update prior interpretations have included updating of interpretive reports (83% of sites), updating of stored annotations (83% of sites), notifying ordering providers about reinterpretation (67% of sites), and notifying patients about reinterpretation (50%).

14% (1/7) of sites have policies in place related to updating prior interpretations as new knowledge emerges.

**EHR with integrated CDS.** 71% (5/7) of sites reported implementing genomics-based clinical decision support within its EHR (both the inpatient and outpatient settings in all cases).

Among these five sites, 100% implemented CDS to recommend treatment based on genetic results, 80% implemented order sets that include genetic testing, 40% implemented CDS to recommend genetic testing triggered by particular orders, and 40% implemented genomics test order templates.

All of these sites implement genomics CDS using their institution's EHRs CDS infrastructure, but one site also employs an external CDS infrastructure.

**Patient portal.** 57% (4/7) of sites report the sharing of genomic reports through a patient portal, while 29% (2/7) report mailing reports to the patient.

**Survey results and the proposed data flow framework**

We believe the survey results validate our proposed data flow framework (figure 1). Participating institutions implement most of the framework components. These institutions also contend with a number of data flow challenges, such as time-consuming manual curation and incomplete discrete data transmission.

The primary framework component that did not appear to be widely implemented, but we believe should remain in the framework, is a systematic process for re-interpreting genomic discrete and pathology reports that remains in step with evolving knowledge. We attribute this finding to the fact that there are currently no trivial automatic methods of performing this task. There is simply too much manual curation required to re-create final reports and interpretative annotations on all previously sequenced patients on a scheduled basis for the life of a patient.

While all sites create final pathology reports and ensure that these results are shared with the ordering providers (in compliance with CLIA requirements), discrete results are less consistently transmitted to downstream systems. This was particularly the case with external laboratory results. Many reports are also stored in PDF format, rather than structured CDA format. Incomplete population of the EHR with both discrete results and structured reports has potential adverse ramifications for CDS. At least some sites rely on entry of variant results into the EHR problem list as a work-around.

We found that all sites automatically generate portions of their draft reports and employ templates. Most sites rely exclusively on their EHR's integrated CDS system to implement CDS.

**Conclusions**

Widespread implementation of precision health will not be feasible in the absence of significant advancements in the automation and standardization of genomic transmission, transformation, and interpretation from instrument to point of care. There is critical need for increased automation of genomic pathology reporting, improved transmission of full discrete variant results and metadata in standardized format to downstream systems, improved methods to store clinically actionable genomic data in EHRs, improved methods to automatically update genomic results for the lifetime of a patient as genomic knowledge evolves, and improved methods to create, maintain, and share comprehensive genomic CDS logic among institutions.

The results of our survey appear to validate our proposed genomic data flow framework. We anticipate that explicit definition of a genomic data flow framework can facilitate discussion of much-needed improvements. Participating survey respondents, arguably leaders in the field of clinical genomics as evidenced by their funding for large scale genomic implementation, are still challenged by current limitations of genomic instruments, bioinformatics pipelines, genomic variant knowledge bases, lab information systems, and EHRs.
Acknowledgements: This work was supported by funding from the NIH-NHGRI grant #U01HG007762 and U01HG010245. We also thank our partners in the IGNITE Network, a consortium of genomic medicine projects funded and guided by the NHGRI (https://ignite-genomics.org/) for their valuable contributions to this project.

Facilitating Analysis of Publicly Available ChIP-Seq Data for Integrative Studies

Avantika R. Diwadkar, MS1, Mengyuan Kan, PhD1, Blanca E. Himes, PhD1
1Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, US

Abstract
ChIP-Seq, a technique that allows for quantification of DNA sequences bound by transcription factors or histones, has been widely used to characterize genome-wide DNA-protein binding at baseline and induced by specific exposures. Integrating results of multiple ChIP-Seq datasets is a convenient approach to identify robust DNA-protein binding sites and determine their cell-type specificity. We developed brocade, a computational pipeline for reproducible analysis of publicly available ChIP-Seq data that creates R markdown reports containing information on datasets downloaded, quality control metrics, and differential binding results. Glucocorticoids are commonly used anti-inflammatory drugs with tissue-specific effects that are not fully understood. We demonstrate the utility of brocade via the analysis of five ChIP-Seq datasets involving glucocorticoid receptor (GR), a transcription factor that mediates glucocorticoid response, to identify cell type-specific and shared GR binding sites across the five cell types. Our results show that brocade facilitates analysis of individual ChIP-Seq datasets and comparative studies involving multiple datasets.

Introduction
Transcription factors play a key role in the regulation of gene expression and the re-organization of chromatin by binding to specific genomic loci in a cell type-specific manner1, 2. Chromatin immunoprecipitation sequencing (ChIP-Seq) is a popular genomic approach to identify binding sites in the genome for a protein of interest3. The technique involves reverse cross-linking of a protein of interest to DNA, followed by fragmentation, immunoprecipitation and parallel sequencing, which permits identification of DNA sequences bound by transcription factors or histones across the genome4. ChIP-Seq has been widely used, with over 40,000 ChIP-Seq assays of various DNA-binding proteins available in public repositories, such as the Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA)5. The availability of these datasets enables researchers to explore novel questions related to transcriptional response by comparing results of multiple ChIP-Seq datasets.

Analysis of ChIP-Seq data is a multi-step process that includes obtaining quality control measures, mapping of sequencing reads to a reference genome, peak calling, performing differential binding analysis, annotating sites, and identifying motifs6. Several specialized informatics tools have been developed to perform these tasks7. The ENCODE project produced many ChIP-Seq datasets along with published standards and guidelines that recommend use of biological replicates, specific antibody and input control characteristics, and uniform processing pipelines to analyze ChIP-Seq data of different protein classes8, 9. For example, to determine whether transcription factor binding sites are associated with corresponding gene transcription, it is recommended that ChIP-Seq experiments include measures of RNA polymerase II (RNAP2) occupancy along with the transcription factor of interest10. In the absence of RNAP2 ChIP-Seq data, genome-wide gene expression results derived from microarray or RNA-Seq studies with similar treatment conditions as the ChIP-Seq study of interest can be used to link transcription factor binding sites with transcriptomic changes.

Software tools to analyze ChIP-Seq data include the R packages chipseq11 and CSAR12, which perform binding site identification based on aligned data obtained via command-line tools, while Homer13, seqMINER14, and Sole-Search15 are popular tools that focus solely on aspects of downstream analysis. Pipelines that combine existing informatics tools to perform comprehensive ChIP-Seq analyses include ChiLin16 and the ENCODE pipeline (https://github.com/ENCODE-DCC/chip-seq-pipeline2), which consist of Python scripts that integrate steps for sequence data mapping, peak calling, and differential binding analysis (only available in ChiLin). The Galaxy-based platforms Cistrome17 and Nebula18 are similar cloud-based options, but their computational speed and customization of commands are comparatively limited. None of the existing pipelines, however, facilitate analysis of publicly available ChIP-Seq data via automated retrieval of raw sequencing files and phenotype information from public
repositories. Thus, there is a need for an end-to-end ChIP-Seq data analysis pipeline that includes direct access to the growing resource of public datasets in a sophisticated yet straightforward manner.

Here, we present brocade, an efficient and customizable ChIP-Seq data analysis pipeline that can be used to analyze publicly available ChIP-Seq data. To demonstrate the utility of brocade, we applied it to study the tissue-specificity of the glucocorticoid receptor (GR). Glucocorticoids are anti-inflammatory drugs commonly used to treat diseases such as asthma\(^1\). At a cellular level, glucocorticoids act by diffusing across the cell membrane and binding to GRs that then translocate to cell nuclei and modulate transcription of various genes in a tissue-dependent fashion\(^2\), including the upregulation of anti-inflammatory genes\(^3\). Although glucocorticoids are known to directly modulate gene transcription via GR binding to glucocorticoid response elements (GREs) and other transcription factors, the cell-specificity of GR-mediated gene transcription is not fully understood. We used brocade to analyze five publicly available GR ChIP-Seq datasets corresponding to five cell types to identify cell type-specific direct transcription targets of the GR.

**Methods**

*Reproducible Analysis of ChIP-Seq Data*

We created brocade, a set of Python scripts that generate bash and R markdown scripts that invoke various functions to facilitate the nearly automated analysis of any ChIP-Seq dataset, with human decision-making required for key steps such as selection of relevant phenotype data (Figure 1). A high-performance computing (HPC) environment is required to run the pipeline efficiently; bash scripts generated by brocade are currently in Platform Load Sharing Facility (LSF) format, a common workload management platform used in IBM clusters. Bash scripts are submitted to an HPC environment to perform GEO/SRA file downloads, read alignment and mapping, peak calling, quality control and differential binding analysis. The data download step can be skipped if a user would like to analyze local ChIP-Seq files. R markdown scripts produced by brocade are used to generate three html reports that (1) describe publicly downloaded datasets, (2) provide quality control metrics, and (3) report differential binding analysis results. Full instructions and code are available on GitHub at [https://github.com/HimesGroup/brocade](https://github.com/HimesGroup/brocade) (individual Python scripts are in the `pipeline_scripts/` directory; R markdown template files are in the `template_files/` directory). We keep separate Python scripts for each of these steps rather than consolidating them into a master script because user input is required at key points during analysis. Additionally, separating steps allows users to more easily understand and customize scripts to suit their needs.

![Figure 1. Steps followed by brocade to analyze ChIP-Seq data along with corresponding R packages (in bold) and other software used.](image-url)

372
Publicly Available Data Download and Phenotype Preparation. Brocade downloads ChIP-Seq raw sequence data in fastq format by retrieving files from an ftp address corresponding to a user-provided GEO accession number using the SRAdb R package. A phenotype file is generated based on information extracted from the phenoData object of the GSE Matrix file that users must confirm is appropriate for desired analyses. Alternatively, users can provide their own fastq and phenotype files.

Read Alignment, Peak Calling, and Quality Control. Adapters, if specified, are trimmed using Trimmomatic22 and overall quality control metrics of fastq files are obtained using FastQC23. Trimmed reads are aligned to the reference genome (currently hg38) using BWA24. Bamtools25 is used to count and summarize the number of mapped reads, while Picard Tools26 is used to compute the insert size of paired-end libraries. Aligned read files are converted to bigwig format, which can be uploaded to the UCSC Genome Browser for peak visualization. DNA-protein binding sites are identified using MACS226 where broad or narrow peak calling is assigned to histone modifications as needed.

Differential Binding and Peak Annotation. Differential binding analysis is performed with the R package DiffBind27, which has built-in functions to retrieve read counts by Rsamtools and GenomicAlignments, and uses DESeq2 to compare binding differences between conditions of interest (e.g., treatment vs. control vehicle)28. The Benjamini-Hochberg method is used to adjust for multiple comparisons and calculate adjusted p-values (q-values). Binding sites are annotated with ChIPseeker29 and binding motif identification is performed with Homer13 and visualized with the R package seqLogo.

Primary GR-Mediated Transcriptomic Changes

ChIP-Seq Datasets. Publicly available ChIP-Seq datasets that measured GR-binding induced by glucocorticoid exposure were obtained from GEO/SRA by searching for the term “glucocorticoid receptor” with Homo sapiens selected as organism. To identify genes with active transcription in/near GR binding sites, we utilized RNAP2 ChIP-Seq datasets provided with GR ChIP-Seq datasets when possible. For GR ChIP-Seq datasets with no available RNAP2 ChIP-Seq profiles, we sought microarray datasets obtained for equivalent cell types under similar treatment conditions in GEO. We utilized brocade to analyze the GR and RNAP2 ChIP-Seq datasets and identify differential binding sites for samples exposed to glucocorticoids vs. control. GR and RNAP2 binding sites were considered to be significantly changed by glucocorticoids if q-values were <0.05. Differential gene expression results for microarray samples were obtained using RAVED (https://github.com/HimesGroup/raved)30. Genes were considered to be significantly differentially expressed with glucocorticoid exposure if q-values were <0.05.

Identification of Shared and Cell Type-Specific Primary GR Target Genes. We defined primary GR target genes as those genes with a GR-binding site within ±20 kb distance of the gene’s transcription start site (TSS), while also having either (1) an RNAP2-binding site within ±3kb of the gene’s TSS (i.e. promoter region), or (2) a significant differential expression result in the corresponding microarray dataset. We identified shared and cell type-specific primary GR target genes by comparing results obtained across all available cell types.

Results

Overview of Brocade Reports

Brocade output html reports include quality control checks and a summary of annotated binding sites for each study of interest. The quality control report has tables and plots illustrating the summary read counts, percentage of mapped and unmapped reads (Figure 2A) and a principal component analysis (PCA) plot based on mapped reads per sample. The differential binding report provides volcano plots (Figure 2B), a PCA plot based on differential binding results, heatmaps, and boxplots of log2 normalized counts (Figure 2C) for significant binding sites identified (i.e., those with q-value <0.05). The report includes select plots from the R package ChIPseeker: frequency of read counts within ±3 kb of the TSS (Figure 2D) and distribution of binding site distance relative to TSS (Figure 2E), as well as sequence logos of top motifs (Figure 2F).

Publicly Available ChIP-Seq Datasets

Our GEO/SRA search yielded glucocorticoid response ChIP-Seq datasets with dexamethasone exposure for five cell types: airway smooth muscle, bronchial epithelial (Beas-2B), adenocarcinomic alveolar basal epithelial (A549), and (2) a differential control, are trimmed using Trimmomatic.
acute lymphoblastic leukemia (RS4;11) and lymphoblastoid (Table 1). RNAP2 ChIP-Seq profiles were available for the three structural cell types (i.e., airway smooth muscle, Beas-2B and A549). We used gene expression microarray datasets GSE71615 (10nM dexamethasone, 24 hr) and GSE44248 (1000nM dexamethasone, 8 hr) for the RS4;11 and lymphoblastoid cells, respectively. Following analysis of ChIP-Seq datasets with brocade, we retained all significant differential binding sites (q-value < 0.05) for the three structural cell types. For the RS4;11 and lymphoblastoid cell datasets, we retained glucocorticoid-induced GR-binding sites with log2-fold change ≥1, as the RS4;11 dataset lacked replicates and no significant GR-binding sites (q-value <0.05) were identified for the lymphoblastoid cell line. Significant genes (q-value <0.05) from the microarray differential expression analysis were also retained.

Table 1. ChIP-Seq datasets selected to study cell-specific GR-binding differences.

<table>
<thead>
<tr>
<th>GEO ID</th>
<th>Cell Type</th>
<th>Control Treatment</th>
<th>Dexamethasone Treatment</th>
<th>GR Control Samples (N)</th>
<th>Treated Samples (N)</th>
<th>RNAP2 Control Samples (N)</th>
<th>Treated Samples (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE95632</td>
<td>Airway smooth muscle</td>
<td>Ethanol, 1 hr</td>
<td>100 nM, 1 hr</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GSE79803</td>
<td>Beas-2B, airway epithelial cells</td>
<td>Ethanol, 1 hr</td>
<td>100 nM, 1 hr</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SRP006756</td>
<td>A549, adenocarcinomic alveolar epithelial cells</td>
<td>Ethanol, 1 hr</td>
<td>100 nM, 1 hr</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GSE71616</td>
<td>RS4;11, acute lymphoblastic leukemia</td>
<td>None</td>
<td>10 nM, 1 hr</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GSE45638</td>
<td>lymphoblastoid cell lines</td>
<td>Ethanol</td>
<td>1000 nM, 1 hr</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

ChIP-Seq Results

We found 39,328 differential GR-binding sites for airway smooth muscle, 20,620 for Beas-2B, 7,818 for A549, 5,113 for RS4;11 and 999 for lymphoblastoid cell line, which corresponded to 6,421, 4,130, 2,297, 1,391, and 137 gene targets, respectively. Most differential GR-binding sites lied beyond 20kb of TSSs (Figure 3A), with a correspondingly large proportion of GR-binding sites located in intronic and distal intergenic regions (Figure 3B). Compared to the large number of GR-binding sites, only 3,220, 1,845, and 2,173 differential RNAP2-binding sites were identified for airway smooth muscle, Beas-2B, and A549, respectively. 8,449 differentially expressed genes for RS4;11, and 262 for lymphoblastoid cell line were identified. Thus, the number of primary GR target genes, based...
on overlap of GR-binding-associated genes with evidence of active transcription (i.e., RNAP2 bound genes or differentially expressed genes obtained via microarray), was 710 for airway smooth muscle, 405 for Beas-2B, 249 for A549, 685 for RS4;11, and 1 for lymphoblastoid cell line (Figure 3C).

**Figure 3.** Characteristics of GR-binding sites for each cell type. (A) Number of GR binding sites within and beyond 20kb of TSS of annotated genes. (B) Proportion of GR binding sites according to type of genomic region. (C) Genes with GR-binding that have (green) vs. do not have (blue) active transcription in response to glucocorticoid exposure as determined by RNAP2 or microarray results. ASM: airway smooth muscle; Beas-2B: airway epithelial cells; A549: adenocarcinomic alveolar epithelial cells; RS4;11: acute lymphoblastic leukemia cells; LCL: lymphoblastoid cell lines.

**Shared and Cell Type-Specific Primary GR Target Genes**

The majority of primary GR target genes were cell type-specific, consistent with our previous transcriptomic findings for genes that were differentially expressed in response to glucocorticoid exposure (Figure 4)\(^{30}\). Because the lymphoblastoid cells only had one primary GR target gene (C16orf87) that did not overlap with the primary GR target genes of any of the other cell types, its results were excluded from Figure 4. Twelve primary GR gene targets, (i.e., FKBPs, MAP3K6, ST3GAL4, IRAK3, NFKBIA, IER2, STX10, ZFP36, TIPARP, IGF2BP3, NBN, DECR1) were shared across airway smooth muscle, Beas-2B, A549 and RS4;11, which included the well-known glucocorticoid-responsive gene FKBPs and IRAK3, a gene that is highly expressed in response to glucocorticoid exposure in airway smooth muscle\(^{36, 37}\) and whose variants have been associated with asthma in European populations\(^{38, 39}\). Thirty-eight additional genes were shared across structural cells, including the known glucocorticoid-responsive genes PER1\(^{36, 37}\), CEBPD\(^{36, 37}\), and DUSP1\(^{36, 37, 40}\). Although Beas-2B and A549 originate from bronchial epithelium, Beas-2B and airway smooth muscle shared more primary GR target genes than Beas-2B and A549 (195 vs. 90).

**Figure 4.** Overlap of primary GR target genes among four cell types.
To confirm shared and cell type-specific results of some primary GR target genes at the level of mapped reads, we checked plots of bigwig files created by brocade. Specifically, we confirmed that there was differential GR-binding induced by glucocorticoids in three cell types for FKBP5 and CRISPLD2, a gene known to be differentially expressed in response to glucocorticoid exposure in airway smooth muscle.\textsuperscript{36, 37} FKBP5 had 4 significant differential GR-binding sites in airway smooth muscle, 6 in Beas-2B, and 6 in A549 (Figure 5A), while CRISPLD2, had 5 in airway smooth muscle, 4 in Beas-2B, and one in A549 (Figure 5B). Differential RNAP2-binding sites for FKBP5 were observed in all cell types (Figure 5A), whereas for CRISPLD2, they were only observed in airway smooth muscle cells after glucocorticoid exposure (Figure 5B).

To confirm that ChIP-Seq results were consistent with transcriptomic results more broadly, we obtained gene expression data for each gene available in our web application REALGAR\textsuperscript{41}, and confirmed that FKBP5 was differentially expressed with glucocorticoid exposure vs. control in all three cell types, while CRISPLD2 was differentially expressed only in airway smooth muscle.

**Figure 5.** Differential GR- and RNAP2-binding sites near genes (A) FKBP5 and (B) CRISPLD2 in airway smooth muscle, Beas-2B, and A549 cells treated with dexamethasone (green) and control vehicle (orange). Asterisk (*) indicates most significant GR or RNAP2 binding site.
Discussion

With the growing interest in leveraging publicly available data, we present brocade, a pipeline that facilitates automated ChIP-Seq data analysis spanning data download from GEO/SRA to differential binding site identification and annotation. As integration of multiple datasets relies on having accurate and consistently obtained results for individual studies, brocade was designed to streamline and standardize ChIP-Seq data analysis. It produces well-documented yet concise html reports that are easily customizable to allow users to include alternative methods. Although brocade is nearly automatic, researchers are required to manually inspect quality control metrics and plots to identify sample outliers and define phenotypes of interest prior to performing differential binding analyses. To maximize use of all publicly available ChIP-Seq datasets, brocade can process standard ChIP-Seq experimental designs that include input DNA and biological replicates to control for genomic artifacts and individual sample variation, respectively1, but also datasets that lack such controls.

There is no consensus on how to assign a specific GR-binding site to a potential gene target on a genomic scale, especially when GR-binding occurs distant to TSS regions. The majority of GR-binding sites have been found to be outside of promoter regions but enriched within 100kb of glucocorticoid-responsive genes42. We restricted our definition of GR-binding sites to include those within 20kb of a target gene’s TSS, a plausible distance that is used in many studies31, 32 and one supported by the fact that among the hundreds of primary target genes we identified, many were well-known glucocorticoid-responsive genes. However, we may have excluded GR bindings that occur in distal enhancers, as the ability to accurately link distal binding events to transcriptional regulation is limited42.

Previous studies have compared protein-DNA binding regions across cell types to identify shared and unique transcription factor binding sites1-43. However, transcription factor binding events alone do not guarantee their role in mediating gene transcription. To identify GR-binding sites that were more likely to mediate glucocorticoid-induced gene transcription, we combined GR-binding site data with evidence of gene transcription provided by RNAP2-binding or differential gene expression results. Previous studies identified glucocorticoid-responsive genes such as FKBP5 across multiple cell types, and PER1 and CRISPLD2 as specific to airway smooth muscle36, 37. Our results found that the transcriptional response of CRISPLD2 to glucocorticoids occurs via promoter-proximal GR binding and is specific to airway smooth muscle, but these cell type-specific findings require further experimental validation. Future studies could compare DNA sequences of the GR-binding sites identified in CRISPLD2 or other genes to determine whether shared and cell type-specific transcriptomic changes involve direct GR binding to GREs or GR-tethering to transcription factors.

Ideally, ChIP-Seq experiments include controls such as RNAP2 ChIP-Seq that help to determine whether transcriptional response accompanies transcription factor binding, but some ChIP-Seq studies use gene expression changes measured by microarray or RNA-Seq as a proxy for changes in transcriptomic activity10. Our results were limited by the design of individual experiments, including different assays used to infer active transcription of genes (i.e., RNAP2 ChIP-Seq for airway smooth muscle, Beas-2B, and A549; microarray for RS4;11 and lymphoblastoid cell line), which likely introduced bias. Other limitations when comparing results across cell types are the heterogeneous dosages of dexamethasone administered, and the fact that one dataset lacked any replicates, which decreases the reliability of its results. Nonetheless, comparison of the GR ChIP-Seq datasets yielded promising results that stand to improve as more publicly available data is generated.

Future versions of brocade will include automated approaches to identify and process publicly available datasets using phenotypes and meta-data extracted from GEO/SRA entries and linked publications to increase its automaticity. For example, for datasets without RNAP2 ChIP-Seq, an automated search through GEO/SRA could identify an appropriately matched gene expression dataset. Future versions of brocade will also include analysis of more recent DNA-protein binding techniques such as ATAC-Seq, DNase-Seq, and STARR-Seq, as these become more widely used.

Conclusion

We developed brocade, a reproducible ChIP-Seq analysis pipeline, that facilitates analysis of publicly available datasets in GEO/SRA. We applied brocade to explore the cell-type specificity of primary GR target genes in five cell types (i.e., airway smooth muscle, Beas-2B, A549, RS4;11, lymphoblastoid). Most primary GR target genes showed cell type-specific activity (e.g., CRISPLD2), but twelve genes were shared across four cell types (e.g., FKBP5). These results indicate potential cell-specific mechanisms of glucocorticoid action that can be explored in future studies.
References

Enabling Privacy Preserving Record Linkage Systems
Using Asymmetric Key Cryptography

Xiao Dong, PhD\textsuperscript{1}, David A. Randolph, MEng\textsuperscript{1}, Subhash Kolar Rajanna, MD\textsuperscript{1}
\textsuperscript{1}Center for Clinical and Translational Science, University of Illinois at Chicago

Abstract
We present a systemic approach to devise and deploy Privacy Preserving Record Linkage (PPRL) systems using asymmetric key cryptography and illustrate the strengths of such an approach. With our approach, the security implications of sharing a common secret salt across the network may be avoided, allowing the local participating sites to use private keys along with the current cryptographic hashes to maximally secure their own data. In addition, the final ciphertext tokens are compatible with those used by existing record linkage modules, allowing seamless integration with the existing PPRL infrastructures for downstream analysis. Finally, study-specific hash production requires action only by the central party. The main intuition for this work is derived from how asymmetric key approaches have enabled internet-scale applications. We demonstrate that such a design, where the local sites no longer need special-purpose software, affords greater flexibility and scalability for large scale multi-site linkage studies.

Introduction
Privacy preserving data sharing for collaborative research across different organizations has recently become a significant challenge in healthcare research\textsuperscript{1}. Data in electronic health records for a single patient may be spread across a number of different healthcare providers. After the patient visits different hospitals and clinics, their electronic health records are registered at the respective data repositories. Harnessing such disjoint information is crucial for establishing a complete longitudinal picture of the patient’s treatment history, assessing the patient’s comprehensive health condition, and improving overall healthcare quality. Complete datasets could be assembled by linking these disparate records; however, doing so directly requires matching certain patient records and may put patient’s privacy at significant risk. Relevant laws and regulations also restrict the sharing of protected health information (PHI) without prior consents from patients. Thus, secure and efficient technologies for privacy preserving record linkage (PPRL)\textsuperscript{2,3,4} are needed to strike the perfect balance between accurate data linkage and privacy protection.

To date, the predominant approach for PPRL is to use cryptographic hashing, such as the SHA-2 family algorithms\textsuperscript{5,6} to conceal PHI. Such a cryptographic hash method is suitable here because it performs a cryptographic transformation, after which the sensitive patient information is converted into hashes—a kind of ciphertext virtually indiscernible from random text strings. Another key property of such hashing algorithms is they work effectively as one-way functions, where the ciphertext (hashes) can be efficiently generated from plaintext but reverting back to plaintext is practically impossible. These two key properties ensure that sensitive identifying information is protected after the hash function is applied. Because the hash function is also deterministic, the same identifying data elements are always transformed to be the same hashes. As a result, the original patient data is de-identified into anonymized tokens and may then be transmitted outside the institution boundary for linkage analysis studies. Such techniques have been deployed in several multi-site network as the principle PPRL approach. Several clinical data research networks (CDRNs)\textsuperscript{7,8,9} funded by the national Patient-Centered Clinical Research Institute (PCORI)\textsuperscript{10,11} serve as prominent examples. Within these research networks, the cryptographic hashes are first generated at different local institutions to de-identify the records, and then submitted to a trusted central party where a designated honest broker performs the tasks of record linkage and deduplication.

Because protecting patient privacy is paramount, it is important to ensure the PPRL systems achieve high security standards. In principle, properly implemented hash functions (such as SHA-256 and SHA-512) constitute a noninvertible function. That is, one cannot reverse engineer the hashing transformation to reveal the identifying data elements. However, a well-known attack against it is the so-called “dictionary attack”, where an attacker uses pre-computed tables to recover the plaintext. For example, if some possible contents are known to be contained in the inputs to the hashing function (such as name, date of birth, etc.), an attacker can construct rainbow lookup tables by producing hashes from the complete set of potentially valid input values. To reduce such security risks, a common approach is to use a “salt,” a random text string added to the original input text. Using a salt renders building the lookup table practically impossible because the attacker does not know what the salt looks like. For PPRL systems relying on cryptographic hash functions, one common salt needs to be used and shared with all participating sites\textsuperscript{12}. This raises security concerns because, if an attacker obtains the salt, a dictionary attack is again possible. Because the salt is common, this security risk applies to all the participants. Therefore, protecting the common salt is essential to
the security of hash-based PPRL systems. It is also worth noting that such security concerns apply whenever a common secret is used, including the “passwords” in PPRL systems using bloom filters.

To meet such challenges, proper security measures for safeguarding the secret salt have been developed. Typically, the secret salt needs to be securely transmitted to the local hosts where the hashing will take place. These hosts have specifications that meet HIPPA security standards, and only authorized employees in the covered entities have access to them. Such an approach becomes unwieldy when the network gets large, however, because all participating sites need first to obtain institutional approval and then securely obtain the hashing application. It becomes especially problematic in a situation where large-scale projects are conducted across networks that adopt different PPRL software products, since now the different vendors need to agree on which common secret salt to use, to disseminate it, and to ensure all parties involved are in compliance with the security protocol. This inevitably leads to technical and policy challenges related to salt distribution and management. Such challenges may vary for different research networks, but they likely always will exist and present significant hurdles for scaling up the PPRL solution across a growing number of networks. As PCORI are planning to carry out large scale and across network linkage, more efficient solutions are needed to mitigate such challenges.

The question then arises: “Can we design an approach to completely eliminate the security implications of sharing a common secret?” In this paper, we describe a novel approach to conduct privacy preserving record linkage using asymmetric key cryptography. With asymmetric keys, the problem of deploying a common secret salt is eliminated because keys are unique for each local institution. Instead of receiving a common salt, local sites instead use their own private keys for encryption and securely transfer their individual public keys to the central party. Critically, the final ciphertext produced remains resistant to cryptanalysis and known attacks. An important additional benefit for the approach is that it allows the PPRL system to scale up more easily, as will be demonstrated below.

Here we highlight the key requirements for such a system. In the method section, we describe how our system, dubbed “PPRL-Plus,” fulfills each of these requirements.

   A. Enable asymmetric key encryption such that local sites use their own secret keys to encrypt their own data, thus eliminating the need for a common secret salt.
   B. Apply strong cryptographic primitives that are sufficiently secure, such as the existing SHA-2 hashing methods and asymmetric key encryption primitives.
   C. Allow easy integration into existing PPRL infrastructures by ensuring outputs remain compatible with existing record linkage tools.
   D. Produce record identifiers in ciphertext that are still resistant to cryptanalysis and known (e.g., dictionary or rainbow table) attacks.

**Method**

In cryptography there are essentially two prevailing paradigms: symmetric key cryptography and asymmetric key cryptography. In symmetric key systems, the same key is used both for encryption and decryption, whereas in asymmetric key systems different keys are applied. One significant advantage for asymmetric key system is only sharing of the public key is required, unlike in symmetric cryptography in which the entire secret key needs to be shared. The common salt approach currently used in PPRL systems is analogous to symmetric key encryption because the same secret (salt) is shared among the local sites in advance, leading to the security concerns and management burdens discussed above. The main focus in our work is to enhance the existing PPRL systems with asymmetric key encryption in order to eliminate the need to use a network wide common secret, while ensuring encryption is sufficiently secure.

The most important applications for asymmetric cryptography include key exchange and digital signatures. Our main intuition derives from digital signatures and given the scope of this paper we are going to skip describing key exchange. In digital signatures, the sender first signs the plaintext message \( m \) using its private key and forwards the encrypted message to the receiver. The receiver then decrypts the message using the sender’s public key in order to verify the authenticity of the message that it actually originated from the sender. One common approach here is to perform a cryptographic transformation first by hashing the plaintext \( m \) and then signing the hash \( h(m) \) using the sender’s private key. The receiver then recovers the hashed message \( h(m) \) using the sender’s public key for verification purposes. In the context of PPRL, \( m \) corresponds to the patient identifying data and \( h(m) \) the hashes produced for record linkage.
Here we briefly go over the well-known RSA algorithm\(^\text{16}\) to illustrate the digital signature process. The first step in RSA algorithm is key generation, where a \{private, public\} key pair is generated as follows:

1. Choose two large primes \( p \) and \( q \) of similar magnitude.
2. Calculate \( N = p \times q \). (\( N \) will be the modulus.)
3. \( \varphi(N) = (p - 1) \times (q - 1) \). (This is the Euler toient function.)
4. Find \( e \), such that \( \gcd(e, N) = 1 \). (\( e \) is the public key.)
5. Find \( d \), such that \( d \times e \equiv 1 \mod \varphi(N) \). (\( d \) is the private key.)

During the signing process, a hashing function is first applied to obtain the hash \( h(m) \) for the plaintext \( m \). The digital signature is then calculated using the private key \( d \) as \[ \text{signature} = (h(m))^d \mod N. \] The signature is sent to the receiver along with \( h(m) \), which is also typically encrypted. The receiver may then verify the signature by comparing the value of \( (\text{signature}^e \mod N) \) to \( h(m) \). (In practice, the hash value \( h(m) \) is subsequently processed according to padding and encoding schemes as specified in standards like PKCS\(^\text{17}\) or OAEP\(^\text{18}\).)

Note that a given signature also constitutes an encrypted (and hashed) message, provided the \( h(m) \) is sent for comparison purposes. But the message is encrypted with a private key and decrypted with a public key, the opposite of what is done in the RSA key exchange algorithm proper. The message is hashed for performance reasons and also to prevent “existential forgeries,” a known security problem with RSA signatures. The receiver is able to verify the signature because the following holds true for properly generated key pairs \{\( e, d \)\} and any non-negative integer \( k \):

\[
(h(m)^d)^e = h(m)^{de} = h(m)^k \varphi(N) + 1 = (h(m)^\varphi(N))^k \times h(m) = h(m)(1)^k = h(m) \mod N
\]  

(1)

Note the digital signature not only uses asymmetric key cryptography but also makes use of a hash function, so it already contains those key components required in our proposed PPRL-Plus system. However, closer scrutiny reveals the last requirement D is not satisfied. The reason is that in the end the hash \( h(m) \) is recovered, unlike in the existing PPRL solutions, where the salt has been mixed in to produce \( h(salt||m) \). Since unsalted hashes are susceptible to dictionary attack, the requirement is not met yet.

The fundamental transformation scheme we want to perform may be illustrated using the following simple example:

Alice: “abc” \( \xrightarrow{\text{PrivateKey}^a} \) “foo” \( \xrightarrow{\text{PublicKey}^a} \) “xyz”  
Bob: “abc” \( \xrightarrow{\text{PrivateKey}^b} \) “bar” \( \xrightarrow{\text{PublicKey}^b} \) “xyz”

where “xyz” \( \neq \) “abc” in both cases.

Here Alice and Bob both hold “abc,” which they each transform into different ciphertext tokens (denoted as “foo” and “bar” respectively) at their local sites using private key encryption. These different tokens are transferred to the central party, where they are subsequently transformed into the same final ciphertext tokens (denoted as “xyz”) using the appropriate public keys. The crucial requirement for this transformation scheme is that reversing the final ciphertext back to the original text must be infeasible (denoted as “xyz” \( \neq \) “abc”). If we apply standard RSA digital signatures, the final tokens will be equal to the original \( h(m) \) value “abc,” which is what is needed. Security concerns arise (in the context of PPRL) if the final ciphertext can be easily reversed back to \( h(m) \), because when \( h(m) \) does not utilize the secret salt, it is susceptible to attack.

To address this, we present a novel method for key generation, based on RSA, to produce asymmetric key pairs that satisfy special mathematical properties. We couple this with a second component: a compression algorithm to produce de-identified tokens suitable for the existing record linkage software modules. More specifically, these ciphertext tokens not only have the same format (such as the commonly used hexdigest format), but also the same cryptographic properties of the regular hashes. We call a ciphertext token containing such properties a pseudo-hash. The first component (Algorithm 1) is executed at the local sites, while the second (Algorithm 2) is executed by the central trusted party to generate tokens for linkage analysis.

The PPRL-Plus key generation algorithm follows:

1. Let \( |N_2| \) be the size of modulus \( N \) in binary. (Default value: 2048.)
2. Let \( |h(m)|_2 \) be the size of the hash in binary. (Default value: 256, when SHA-256 in use.)
3. Choose \( 1 < r < \frac{|N_2|}{|h(m)|_2} \). (Default value: 7.)
4. Execute steps 1 through 3 of RSA Key Generation.
5. Find the key pair \( \{d', e'\} \), where \( d' \times e' = r \mod \varphi(N) \).
6. \( d' \) is the private key, and \( e' \) is the “protected” key.

**Algorithm 1.** Modified key generation for PPRL-Plus. The subscripts “2” here denote the binary format.

In this algorithm, we introduce a new integer parameter \( r \), such that the binary width of \( h(m)^r \) is guaranteed to be less than that of \( N \) (the same as RSA key size). This holds when the size of \( h(m)^r \) in binary is at least \(|h(m)|_2\) bits less than that of \( N \) in binary. For example, in the algorithm above where the default SHA-256 hashing function is used, the hashes \( h(m) \) will be exactly 256 bits, using the default value of \( 7 \) for \( r \), \( h(m)^r \) will be at most \( 256 \times 7 = 1792 \) bits wide, which is guaranteed to be smaller than the modulus \( N \), a 2048-bit integer. Hence, the following holds true.

\[
(h(m)^{d'} e')^e = h(m)^{d' \times e'} = h(m)^{\varphi(N) \times k + r} = (h(m)^{\varphi(N)})^k \times h(m)^r = h(m)^r \mod N
\]

(2)

The above transformation corresponds to the original RSA verification algorithm. The main difference is that with the set of key pairs \( \{d', e'\} \), the original hash \( h(m) \) is transformed into \( h(m)^r \). The signing produces hash signatures \( h(m)^d \mod N \) and it takes place at the sender just as in the original digital signature procedure. Using properly constructed one-way compression function \( \gamma(h) \), it can be subsequently converted into a pseudo-hash \( h(m)^r \) that satisfy requirement D, as the following:

\[
\gamma(h(m)^r) = h(m)', \text{ where } h(m)' \neq h(m) \text{ and } |h(m)|_2 = |h(m)|_2.
\]

(3)

Here, the following key property must hold for \( \gamma(h) \): it takes an input whose binary size is approximately \( r \) times the binary size of the original hash \( h(m) \), and outputs a pseudo-hash whose binary size is exactly the same as \( h(m) \). For example, in our default setting where the original hash is produced by SHA-256 and \( r = 7 \), \( \gamma(h) \) takes input that is approximately 1792 bits long and outputs a pseudo-hash that is exactly 256 bits long, which is conveniently the same width as a SHA-256 hash value to be consumed by the existing hash matching modules. In most PPRL applications, the hashes are processed into the corresponding hexadecimal format using a digest function. For example, if SHA-256 is used, the outputs will be strings contain 64 hexadecimal characters. In our implementation, we take a number of randomly selected 256-bit blocks from \( h(m)^r \) and iteratively apply them as XOR stream ciphers. Due to the pseudo-randomness of the hash function in \( h(m) \), each 256-bit blocks randomly selected from the binary form of \( h(m)^r \) can be treated as a pseudo-random number, hence as a stream cipher. More formally, the compression algorithm can be described as follows:

<table>
<thead>
<tr>
<th><strong>Input</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>( h(m)</td>
</tr>
<tr>
<td>( s : ) seed for random number generator (RNG)</td>
</tr>
<tr>
<td>( k : ) number of times XOR stream ciphers applied</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Output</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>( h(m)}_{16} : ) pseudo-hash in hexadecimal format</td>
</tr>
</tbody>
</table>

\[
\text{RNG} \leftarrow s \\
\text{index} \leftarrow \text{RNG}(|h(m)|_2) \\
\text{size} \leftarrow |h(m)|_2 \\
\text{block} \leftarrow h(m)\}_{index: \text{index} + \text{size}, \text{circular}} \\
\text{repeat } k \text{ times:} \\
\text{index} \leftarrow \text{RNG}(|h(m)|_2) \\
\text{stream} \leftarrow h(m)\}_{index: \text{index} + \text{size}, \text{circular}} \\
\text{block} \leftarrow \text{block} \oplus \text{stream} \\
\text{h(m)}_{16} \leftarrow \text{hexdigest(block)}
\]

**Algorithm 2.** Pseudo-hash compression algorithm. The subscripts denote binary or hexadecimal formats.

In this algorithm, the block size is defined as the standard hash size in binary. For example, if SHA-256 is used, the block size will be 256. The starting block index is randomly selected by a random number generator (RNG) from a range of integers up to the binary width of \( h(m)^r \), and indices are implemented as a circular array, such that the entire
range of \( h(m) \) may be utilized. Using a fixed value for the seed, the sequence of integers generated for the block indices are guaranteed to be the same every time the above algorithm runs. The seed is an important security parameter and must be kept secure. After the initial block is selected, the XOR stream cipher is repeatedly applied \( k \) times. Note that \( k \) must be chosen carefully. To satisfy requirement D, it must be infeasible to recover \( h(m) \) via \( h(m)^r \). It is crucial to ensure cryptanalysis of pseudo-hash output \( h(m)^r \) generated via Algorithm 2 cannot reveal \( h(m)^r \). Fortunately, without prior knowledge of how random indices are chosen, the probability of finding all correct stream ciphers can be estimated as follows (assuming \( r = 7, \) and \( |h(m)|_2 = 256 \) for SHA-256):

\[
\prod_{i=1}^{k} \frac{1}{|h(m)|_2} \leq \left( \frac{1}{256 \times (r-1)} \right)^k = \left( \frac{1}{2^{8 \times \log(20)}} \right)^k < \left( \frac{1}{2^{10.5}} \right)^k < \frac{1}{2^{512}} \quad \text{when } k = 50.
\]

Therefore, when we choose \( k \) appropriately, such probability is even smaller than finding a collision for a SHA-512 hash. In the above equation, where \( k \) is set to 50, an intruder has no reasonable way to invert the pseudo-hash output \( h(m)^r \) back to \( h(m) \), and then proceed to recover \( h(m) \). As a general guideline, \( k \) should be determined by the parameter \( r \) and the block size \( |h(m)|_2 \). Using such \( k \), enough random blocks are sampled from \( h(m)^r \) to ensure the chance that any parts of it were never passed through the XOR stream cipher would be very low. Specifically, using our default values \( (r = 7, k = 50 \text{ and SHA-256}), \) the probability that a given bit is never chosen is \( \prod_{i=1}^{k} P_i \cong \left( \frac{1}{2^{10.5}} \right)^k < 0.05\% \). Note that approximately 6 out of 7 bits within \( h(m)^r \) won’t be chosen as a part of the XOR stream cipher for each round.

Here we complete the discussions for the main algorithmic methods in PPRL-Plus system. The new key generation algorithm is a slight modification based on the widely used open standards of RSA key generation, and it not only offers the local sites the preferred private keys for encryption, but also does not require use of a common secret salt. Compared to existing PPRL systems in which all the participating parties needs to have special software to retrieve the salt and to produce the hashes, only the central trusted party needs one piece of additional software to run the pseudo-hash compression algorithm. Because it does not require the efforts to coordinate and disseminate the salt, this promises to be a more scalable solution to incorporate a large number of participating sites to perform intra-network record linkage studies. In the next section, we discuss how we have operationalized the different algorithmic components in a research data network setting, where we simulate a multi-site study. Notice such mode of operations also applies in multi-site studies that span several different research networks.

**Results**

In this section, we first explain how to deploy the key generation algorithm (Algorithm 1). The default value for \( r \) in the original RSA key generation algorithm should be 1 as always, the PPRL-Plus key generation algorithm can be achieved by using a new value for \( r \) in the product of \( d \) and \( e \mod \varphi(N) \) (as shown in step 5 of Algorithm 1). Using a good quality open source project, such an algorithm can be implemented quite easily. In our experiment, we set the default values for \( r \) to 7 as demonstrated above. New key pairs were generated around one second on desktop hardware.

![Figure 1. Key generation and hash signature generation workflow.](image-url)
Notice the key pairs in our discussion here are only meant to be used in the PPRL applications to encrypt the hashes. They are not the same as the regular RSA keys of the host server. To differentiate these keys, we call them PPRL-Plus keys. Once the key pair is generated, the private key \( d \) is used to encrypt the hashes created from the cryptographic hash algorithms (such as SHA-512 or SHA-256). This step corresponds to generating \( h(m)^d \mod N \). Here we leverage the security primitive that finding the modular root of an arbitrary number (hash signature) is infeasible when the key size is large enough. In our implementation the key size of 2048 bits is used according to the NIST recommendation\(^\text{19}\).

Notice the secret salt is no longer required when creating the hashes because the security for the hashes is now protected by the private key to which only the local site has access. The resulting hash signatures not only apply a cryptographic hash algorithm to de-identify the PHI, but also are sufficiently protected by the site’s own private key. Hence, they can be transmitted outside the institutional boundary to a file repository hosted by a central party for downstream analysis.

One major issue that needs to be addressed here is that we need to enhance the security measures with respect to the supposedly public key component \( e \). In regular RSA key generation, the public key \( e \) may be shared with impunity. This is expressly not the case for the method we describe here. If somehow an intruder obtains the private hash signatures \( h(m)^d \mod N \), the known value of \( e \) can be used to derive \( h(m)^e \), which can be mapped to \( h(m) \) for dictionary attacks. We thus call the \( e \) component a protected key in the PPRL-Plus system, as compared to public key in the regular RSA application. The resulting key \( e \) is then encoded in PEM format and securely forwarded to a central party when requested, as indicated in Figure 1.

![Figure 1. Pseudo-hash generation workflow.](image)

The pseudo-hash compression and the subsequent data distribution workflow is illustrated in Figure 2. In order to produce hashes suitable for linkage, the following actions are performed. A signature repository is populated with encrypted data from the various local institutions. The trusted party (also referred to as the honest broker) uses the pseudo-hash compression engine to process the data received from the various local institutions. As needed, protected keys are retrieved from the local sites and the hash signature files from signature repositories. The honest broker plays a coordinating role throughout this execution phase. During execution, the engine picks a fixed seed to ensure the XOR stream ciphers are obtained from the same set of blocks within \( h(m)^y \). Therefore, the same \( h(m)^e \) values from different local institutions are be compressed to the same pseudo-hashes \( h(m)^e \) for the down-stream linkage. During execution, details that have significant security implications (namely, the seed and the protected keys) reside only in the host machine’s memory. These data are never written to disk. Furthermore, they are properly scrubbed from
memory at the earliest opportunity. Notice that the private hash signatures $h(m)^d \ mod \ N$ by itself is sufficiently secure as long as the private key $d$ meets the standard size requirement.

Since all the submitted data needs to be processed by the compression engine, it is necessary that the compression algorithm offers acceptable speed performance. In our experiment, we implement the pseudo-hash compression algorithm using Python version 3.7.3, and we observe a simulated average speed of 0.4 millisecond per record using an Intel i7 8th generation processor. This translates to approximately 7 minutes to generate one million pseudo-hashes using a single-threaded application. Given this performance, we are convinced the compression algorithm is efficient enough to generate study-specific pseudo-hashes for multiple studies. On the other hand, we observe the average simulated speed to generate the private hash signatures is about 21 milliseconds, which translates to approximately 6 hours for 1 million private hash signatures. This is primarily due to the security requirement to use a sufficiently large private key. While such performance speed appears slow, it might actually be considered acceptable in practice, because the private hash signatures need only be generated once at each local site. In addition, such a performance issue may be addressed easily using parallel processing or hardware acceleration.

In large-scale data research networks where multiple studies are carried out, it is common to create hashes that may only be used for individual projects. This is necessary to comply with various data governance policies. In current PPRL systems, hashes created for one particular study may not match with hashes created for other protocols. Therefore, study-specific hashes must be generated using different secret salts, and all local sites need to run the hashing application as many times as there are studies requiring their site-specific hashes. This becomes unwieldy as the scale of the data research network grows larger, and more studies are in demand. Note that in PPRL-Plus, if a different seed is used in the compression algorithm, a different set of XOR stream ciphers will be obtained, and different pseudo-hashes will be generated. However, the pseudo-hashes created even with a different seed will still match if they originate from the same $m$ and $h(m)$. Such characteristics ensure pseudo-hashes are only capable of linking records under one study and cannot be used for linkage to other studies, assuming a different seed is used for each study. Moreover, since pseudo-hash generation is performed at an outside entity, the local institutions no longer need to re-generate linkage hash tokens for different studies. Thus, PPRL-Plus provides better scalability for conducting multiple concurrent projects. The resulting study-specific pseudo-hash tokens are further transferred to different study teams for downstream analysis, as illustrated in Figure 2.

**Discussion**

In the following table we compare our proposed PPRL-Plus system to the prevailing PPRL security measures. As we have explained previously, the main advantage for PPRL-Plus is to achieve the goal of using site-specific private keys for encryption. In order to produce the site-specific private hash signatures $h(m)^d \ mod \ N$, first hashes are calculated using the standard SHA-2 family functions. The hashing functions required such as SHA-256 or SHA-512 come with modern programming language and database applications and can be easily applied. Second, only slight modification to the open source RSA key generation library is required to change the remainder value from the default 1 to $r$. Third, the private hash signatures $h(m)^d \ mod \ N$ may be calculated efficiently. Fast implementations for the modular exponentiation function required in the third step are also available, such as the $\text{pow()}$ function from Python’s standard build-in library. Therefore, it is straightforward for participating sites to adopt this approach locally to create private hash signatures without using special-purpose vendor software.

**Table 1.** Comparison of PPRL-Plus and current PPRL systems.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>PPRL-Plus</th>
<th>PPRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Require distribution of network wide common salt/secret</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Need special-purpose software at local sites</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Need special-purpose software at the central</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Encryption speed at local sites</td>
<td>Slow</td>
<td>Fast</td>
</tr>
<tr>
<td>Encryption speed at the central</td>
<td>Fast</td>
<td>N/A</td>
</tr>
<tr>
<td>Generate study specific hashes at local sites</td>
<td>One time</td>
<td>Many time</td>
</tr>
</tbody>
</table>

Note that in regular digital signatures, padding and encoding schemes such as PKCS or OAEP are applied to prevent malleability attacks. However, in the context of PPRL, the main security threat is not about altering ciphertext.
Therefore, we may skip such padding schemes. If such padding is required, it is necessary to use a smaller value for the remainder \( r \), such that the conditions in equation (2) in the method section still hold. The resulting private hash signature is secure in its own right and may be transmitted securely outside the institutional boundary, as long as the key size is sufficiently large. Currently the NIST recommendation for the key size is 2048 bits. Guaranteed by the demonstrated security of the RSA formulation, the hash signatures \( h(m)^d \ mod \ N \) saved in the signature repository by themselves are resistant to cryptanalysis. Armed only with these signatures, intruders have no way to recover the original \( h(m) \). If a private hash signature were exposed, the relevant local site need only produce a new key pair, regenerate the private hash signatures, and permanently delete the old key pairs. Following such a recovery procedure, normal operations may be resumed. Because different local sites use different key pairs, the impact on any individual security instance is confined to a particular local institution. On the contrary, in common salt-based PPRL systems, the compromise of any individual institution may have broader impact.

Also note that while executing the pseudo-hash compression algorithm, the value of \( h(m)^r \) for the unsalted hash \( h(m) \) is present in the memory of the host machine. Since \( r \) is employed as a common public parameter, one can easily replace \( h(m) \) with \( h(m)^r \) in a dictionary table. Therefore, carrying out a cryptanalysis attack using \( h(m)^r \) is just as easy as using \( h(m) \). The seed for the random number generator also persists in memory, albeit for short period of time, posing another potential risk by possibly revealing how the XOR stream ciphers are applied. It is therefore very important to ensure the security of the host machine where the pseudo-hash compression engine runs, and to flush the memory completely as soon as these sensitive data are no longer required.

To eliminate such vulnerability completely, we seek a method to achieve the fundamental transformation scheme presented in page 3 without using resorting to intermediary processing that may reveal the original hash \( h(m) \) or any of its derivable forms. In future work, we will be seeking a function with suitable mathematical properties that can directly transform the private hash signatures into pseudo-hash tokens that are not practically reversible to the original hash values.

**Conclusion**

In this paper, we have demonstrated a novel framework to carry out privacy preserving record linkage for multi-site studies using asymmetric key cryptography. The main advantage of our approach is greater scalability by eliminating the need for pre-disseminating a common secret to participating sites. Our inspiration is derived from how asymmetric keys have helped scale up secure Internet applications. We have provided an overarching demonstration of how to design, implement, and deploy a PPRL system based on asymmetric keys. This work is still in its early stages, and we are working to identify functions with better mathematical properties to meet the fundamental transformation requirements.

**References**

Machine Learning Based Opioid Overdose Prediction Using Electronic Health Records

Xinyu Dong, Sina Rashidian, Yu Wang, Janos Hajagos, Ph.D., Xia Zhao, Richard N. Rosenthal, MD, Jun Kong, Ph.D., Mary Saltz, MD, Joel Saltz, Ph.D., MD, Fusheng Wang, Ph.D.

Stony Brook University, Stony Brook, NY

Abstract

Opioid addiction in the United States has come to national attention as opioid overdose (OD) related deaths have risen at alarming rates. Combating opioid epidemic becomes a high priority for not only governments but also healthcare providers. This depends on critical knowledge to understand the risk of opioid overdose of patients. In this paper, we present our work on building machine learning based prediction models to predict opioid overdose of patients based on the history of patients’ electronic health records (EHR). We performed two studies using New York State claims data (SPARCS) with 440,000 patients and Cerner’s Health Facts database with 110,000 patients. Our experiments demonstrated that EHR based prediction can achieve best recall with random forest method (precision: 95.3%, recall: 85.7%, F1 score: 90.3%), best precision with deep learning (precision: 99.2%, recall: 77.8%, F1 score: 87.2%). We also discovered that clinical events are among critical features for the predictions.

Introduction

The United States is experiencing an epidemic of opioid related deaths due to the misuse and abuse of opioids, including both prescribed pain relievers and illegal drugs such as heroin and synthetic fentanyl. According to Han1, 91.8 million (37.8%) civilian non-institutionalized adults in the US used prescription opioids in 2015. Among them, 11.5 million (4.7%) misused them and 1.9 million (0.8%) had a use disorder. Between 2005 and 2014 OD-related hospitalizations and emergency room visits increased 64% and 99% respectively.2 According to Scholl L3, drug overdoses resulted in 70,237 deaths during 2017, and 47,600 (67.8%) of them involved opioids.

Improving current clinical practices can potentially reduce the risks of opioid misuse or overdose. For example, CDC4 has provided recommendations for prescribing opioids for chronic pain care on opioid prescription for primary care clinicians. The joint commission also provided new and revised pain assessment and management standards for accredited hospitals. Clinical decision support systems (CDSS) can provide medical advice through integrating a patient’s EHR data for inference, which holds high potential for improving clinical decisions and clinical practices. For instance, using EHR data to help physicians provide better care in opioid use disorder is proposed5, which depends on understanding the past history of patients and using pre-defined rules for the decision support. It is particularly important to predict and identify patients at risk for opioid toxicity in time for optimal clinical interventions, such as reducing opioid dosage or suggesting alternative options for chronic pain management.

CDSS is made possible with the adoption of EHR, which has increased dramatically with the introduction of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009. Besides EHR data managed by healthcare providers, large scale EHR data are also made available through government open data initiatives, such as claims data from the New York State Statewide Planning and Research Cooperative System (SPARCS).7 In addition, commercial EHR vendors often provide large scale de-identified EHR data for research purposes; for example, Cerner’s Health Facts is a large multi-institutional de-identified database derived from EHRs and administrative systems.

Traditional machine learning has been widely explored and discussed in previous work8,9,10,11. Recently, deep learning methods are gaining popularity in EHR based predictive modeling. For instance, Rajkomar et al. performed a large scale deep learning-based study with high prediction accuracy using EHR data in multiple medical events prediction12. Another study employed a fully connected deep neural network to suggest candidates for palliative care.13 Recurrent neural networks (RNN) are also applied to take advantage of patient’s background information for prediction of clinical events12,14,15,16. For example, one study explored the application of RNN for chronic disease prediction using medical notes17. Our recent work has applied fully connected networks for predicting diseases and improving coding18,19. For opioid related application, Che20 followed a deep learning based method to classify opioid patients into opioid-
dependent and long-term users. Ellis used machine learning classifiers to predict the likelihood of patients having substance dependence 21.

In this paper, we built multiple prediction models for predicting the risk of opioid poisoning in the future using patients’ history from claims data and EHR data respectively, and examined most important features for such predictions. We leveraged large scale databases to identify features that are commonly associated with opioid poisoning, while minimizing irrelevant features. We used the medical records of a patient for the prediction, including demographic information and past medical history including diagnoses, laboratory results, medications and related clinical events. Both traditional machine learning algorithms and deep learning methods were studied in our research. Our results demonstrated that with comprehensive EHR data such as Health Facts, our models provided highly promising prediction results: we were able to generate a best recall using Random Forest based method (precision: 95.3%, recall: 85.7%, F1: 90.3%, AUC: 95.1%), and best precision and Area Under the ROC Curve (AUC) with Neural Networks (precision: 99.2%, recall: 77.8%, F1: 87.2%, AUC: 95.4%). We also discovered that clinical event features play critical roles in the prediction.

Methods

Data Sources

We extracted inpatient EHR data from two clinical databases, SPARCS and Health Facts, as our data sources. 

SPARCS Inpatient Data. We used hospital discharge data from New York State SPARCS7 database. New York State requires any New York State healthcare facility (Article 28 licensed) certified to provide inpatient services, ambulatory surgery services, emergency department services or outpatient services to submit data to SPARCS. The purpose of SPARCS is to create a statewide dataset for providing high quality medical care by serving as an information source.

Cerner’s Health Facts Inpatient Data. We also utilized Cerner’s Health Facts as another data source for our studies. Health Facts includes de-identified EHR data from over 600 participating Cerner client hospitals and clinics in the United States. In addition to encounters, diagnoses, procedures and patients’ demographics that are typically available in claims data, Health Facts also includes medication dosage and administration information, vital signs, laboratory test orders and results, surgical case information, other clinical observations, and health systems attributes22.

Diagnosis Codes for Data Extraction

All patients were classified into two groups, namely opioid overdose patients and other patients. Opioid overdose patients are defined as those who received at least one opioid related diagnosis code in their medical records. For those patients who have not been diagnosed with those codes, we took them as negative (non-opioid poisoned) patients.

The opioid poisoning related diagnosis codes are a collection of ICD-9 codes and ICD-10 codes (starting from October 1, 2015) for poisonings by opiates, opium, heroin, methadone, and other related narcotics. According to Moore23, the selected ICD-9 codes include 965.0 (Poisoning; Opiates and Related Narcotics), 965.00 (Poisoning; Opium/alkaloids, unspecified), 965.01 (Poisoning; Heroin), 965.02 (Poisoning; Methadone), 965.09 (Poisoning; Other opiates and related narcotics), E85.00 (Accidental Poisoning; Heroin), E85.01 (Accidental Poisoning; Methadone), 97.01 (Poisoning; Opiate antagonists) and E85.02 (Accidental Poisoning; Other Opiates and Related Narcotics), selected ICD-10 codes include T40.4 (Poisoning; other synthetic narcotics), T40.0 (Poisoning; opium), T40.1 (Poisoning; heroin), T40.2 (Poisoning; other opioids), T40.3 (Poisoning; methadone), T40.6 (Poisoning; narcotics), all of the codes and their descendant codes.

Although we don’t know if those negative patients will get opioid overdose in the future, we can take their status of opioid poisoning in the last visit as the future we want to predict. We can say that they have low risk of opioid poisoning in near future. Then we can use all the information of visits before last one as features for prediction.

Study Datasets

We built two independent study datasets for SPARCS and Health Facts respectively. We extracted records from January 13, 2005 to December 25, 2016 from SPARCS. We selected patients with at least one historic encounter before first opioid poisoned related diagnosis. We randomly chose 40,000 positive (opioid poisoned) patients. As the dataset would be highly imbalanced by having all non-opioid related patients, we chose 400,000 negative patients without a history of opioid poisoning.

For Health Facts database, we extracted records from January 8, 2000 to December 29, 2017. The selection of patients is similar to SPARCS data, but we further filtered patients and retained those who had at least one hospital visit with
a clinical event record to help in evaluating the importance of clinical events. We used 110,000 patients from Health Facts, with 10,000 positive ones and 100,000 negative ones.

**Features.** Information useful to predict future opioid overdose includes diagnosis codes, procedure codes, medications, clinical events and demographic information. Since SPARCS and Health Facts are very different datasets, the features from each dataset vary significantly.

Diagnosis codes specify diseases, symptoms, poisoning for patients, and the history of diseases are critical information for predicting the future. Due to the large space of ICD codes, in order to prevent influence of too many unrelated or biased features, we filtered the diagnosis codes based on their frequency of co-occurrence with opioid poisoning, or their frequency appeared in the history of opioid poisoned patients. For reliability, we removed all opioid poisoning related diagnosis codes to prevent the model from directly getting results from the related codes. Diagnosis codes are used for both SPARCS and Health Facts datasets.

Procedure codes are specific surgical, medical or diagnostic interventions received by patients. Procedure codes are extracted for both datasets.

Medications are recorded by NDC codes, which is available in Health Facts dataset only. Preprocessing medication codes followed the same procedures as diagnosis codes and procedure codes. After filtering, only 10% of them are selected in the feature space. As dosage and duration of medications can be important, they are added to the features space using medication codes. For a specific medication, we measured the time period that the patient took it and the total dosage the patient took during the period, and had them added to the feature space. Quantity of dosage is measured in according unit in the database, to illustrate it, liquid dosage such as Toradol IV/VM (ketorolac) is measured in tubes, solid dosage such as Pepcid (famotidine) is measured in tablets, and duration is measured in minutes.

Clinical events are related symptoms, procedures, and personal situations that are not formally classified into any codes above, which is available in Health Facts dataset only. Some of them are identified as clinical events in Health Facts, for instance, the pain level of patients, smoke history, height, weight, and travel information. Since 79.21% of hospitals in Health Facts have clinical event records, they can be helpful for most hospitals for prediction.

Demographic information such as age, gender and race are added to the feature space as well to improve the prediction. Since they are relevant to a patient’s social-economic status, we included it in the feature space in SPARCS dataset.

**Feature Selection and Normalization**

While there is a large number of codes from diagnosis, procedures and medications, many of them are not useful for prediction. Gathering all codes would also create a gigantic feature space that would then result in very slow training process. Therefore, we filtered the features to compress the feature space. For each feature category, we ranked all codes by their frequencies of occurrence in the history of patients with opioid poisoning. We kept the top 10% features in each category in the feature space. Note that Health Facts has more comprehensive features than SPARCS. The numbers of features selected are summarized in Table 1.

**Table 1. Summary of selected features for prediction in study datasets.**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Category</th>
<th># of Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPARCS</td>
<td>Diagnosis</td>
<td>2000</td>
<td>ICD-9 and ICD-10 codes</td>
</tr>
<tr>
<td></td>
<td>Procedure</td>
<td>2000</td>
<td>ICD and CPT codes</td>
</tr>
<tr>
<td></td>
<td>Demographic information</td>
<td>4</td>
<td>Race, gender, age and payment method</td>
</tr>
<tr>
<td>Health Facts</td>
<td>Diagnosis</td>
<td>2000</td>
<td>ICD-9 and ICD-10 codes</td>
</tr>
<tr>
<td></td>
<td>Procedure</td>
<td>1000</td>
<td>ICD and CPT codes</td>
</tr>
<tr>
<td></td>
<td>Demographic information</td>
<td>3</td>
<td>Race, gender and age</td>
</tr>
<tr>
<td></td>
<td>Clinical Events</td>
<td>900</td>
<td>500 events and numeric value for 400 of them</td>
</tr>
<tr>
<td></td>
<td>Medication</td>
<td>4500</td>
<td>1500 NDC codes with does quantity and duration</td>
</tr>
</tbody>
</table>
For opioid poisoning patients, we used features from visits before they were diagnosed with opioid poisoning for the first time. We took advantage of all visits before the last one for non-opioid poisoning patients, features in all of those visits will be used for prediction. Their status of opioid poisoning in the last visit is the target we want to predict.

For ages, we segmented them into multiple age groups according to the ages of last visit: the first age group is 0-5, followed by every 10 years. This can accelerate the training process with minimal impact on the performance. For categorical features, for example, race and payment methods, One-Hot encoding method was applied to encode those features. Figure 1 illustrates how One-Hot Encoding is implemented for sources of payments. For Patient 1, a code [1,1,0,0,0,0,0] will be generated.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Self-Pay</th>
<th>Medicare</th>
<th>Worker’s Compensation</th>
<th>Medicaid</th>
<th>Insurance Company</th>
<th>Blue-Cross</th>
<th>Other Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Patient 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Patient 3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. An Illustration of the One-Hot Encoding on Sources of Payment.

We used a binary representation for diagnosis codes, procedure codes, medication codes and clinical events. If one of these features is detected from a patient’s history, the feature value is 1, otherwise the value is 0. Other features, such as medication dosage quantity and duration, clinical events of numeric values such as blood pressure, height, pain score, are assigned with numeric value and are measured in according units in the database, for instance, height measured in centimeters, blood pressure measured in mm[Hg]. The feature preprocessing is illustrated in Figure 2.

Figure 2. Example of feature preprocessing or the prediction models.

**Prediction Methods**

The goal of this study is to predict opioid poisoning of patients in the future using existing EHR data. We built our prediction models with multiple machine learning methods. Traditional machine learning methods such as decision tree and random forest have been proved effective in many health data analytics applications. Recently, deep learning methods have been widely used due to the capability of handling large number of features. Since there are a large number of features in our studies, we decided to take advantage of deep learning methods as well and compared that with traditional methods.

**Traditional Machine Learning Methods.** We utilized methods including random forest, decision tree and logistic regression in our work, which have proven effective for EHR data-based prediction tasks. In the experiments,
we kept most settings and parameters with default values, and adjusted some of them to fit our tasks. In practice, for clinical decision support, identifying all patients with high opioid poisoning risk is desired, thus between high precision and recall rates, the high recall rate in the prediction model can be more critical. In logistic regression, we employed l2 regularization as the penalization and liblinear as the optimizing algorithm. Both random forest and decision tree models took Gini impurity as the criterion to split data in training, which will be further discussed for feature importance calculation below. Furthermore, there is no tree depth limit for either random forest or decision tree model. We assigned a 10 times higher weight to positive cases than negative cases for all three methods during the training phase in order to achieve a higher F1 score, which can potentially solve the problem that the dataset is imbalanced.

**Deep Learning Methods.** Deep neural networks have been proven effective in many healthcare prediction applications\(^\text{13, 14, 16, 20, 27}\). We implemented a fully connected neural network based model. Our model is composed of three fully connected layers, two dropout layers and one output layer. First two fully connected layers are each followed by a dropout layer and then followed again by one more fully connected layer, which is connected to the output layer. We set the dimension of first two fully connected layers as 512 and third fully connected layer as 8. Since this is a binary classification task, the dimension of the last layer is one. A dropout layer follows a fully connected layer to prevent overfit - it will randomly drop out a portion of outputs from previous fully connected layer at each training epoch. We set the portion or namely dropout rate as 20%. The framework of our neural network is illustrated in Figure 3.

![Figure 3. The framework of the neural network designed for the prediction model.](image)

A rectified linear unit (ReLU) function is used as the non-linear transformation function for each hidden layer. We chose ReLU for preventing vanishing gradient and sparse activation problems, and improving computation efficiency\(^\text{28}\). A sigmoid function is applied in the last layer. We learned the parameters by optimizing binary cross-entropy loss function. To minimize the loss function, Adam optimization algorithm was selected. Adam is an optimization algorithm can replace classical stochastic gradient descent procedure to update network weights iterative, it has advantages in computation efficiency and little memory requirement for it combines the best property of adaptive gradient algorithm and root mean square propagation\(^\text{29}\).

**Implementation.** In our experiment, the implementation environment is the Python programming language (2.7). Traditional machine learning methods are implemented with Python scikit-learn package\(^\text{30}\). Deep learning is implemented with Python Tensorflow\(^\text{31}\) and Python Keras\(^\text{32}\). Other used libraries include Python Numpy and Python Pandas. The training was performed on an NVIDIA Tesla V100 (16GB RAM).

**Results**

**Prediction Results Analysis**

In our experiments, for both SPARCS and Health Facts datasets, we randomly took 90% of positive and negative patients as training set and rest as test set. To be specific, 360,000 negative patients and 36,000 positive patients as training set, while 40,000 negative patients and 4,000 positive patients as test set, for SPARCS dataset. 90,000 negative patients and 9,000 positive patients as training set, while 10,000 negative patients and 1,000 positive patients as test set, for Health Facts dataset.
To comprehensively evaluate performance of models, we calculated all common metrics including accuracy, precision, recall, F1 score and AUC. As the dataset is imbalanced, accuracy and AUC can be misleading. As we stated in the previous section, recall is a critical factor for the prediction models. The F1 score is a measurement considering both precision and recall, so it is an aggregated assessment of the overall prediction performance. We compared traditional methods and deep learning method for two datasets respectively. The results are shown in Table 2. The best results for each metric category are highlighted in bold.

**Table 2.** Experiment Results of different methods on SPARCS and Health Facts datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prediction Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPARCS</td>
<td>Random Forest</td>
<td>96.03%</td>
<td>88.96%</td>
<td>64.25%</td>
<td>74.61%</td>
<td>94.94%</td>
</tr>
<tr>
<td></td>
<td>Decision Tree</td>
<td>95.35%</td>
<td>75.87%</td>
<td>71.68%</td>
<td>73.71%</td>
<td>84.62%</td>
</tr>
<tr>
<td></td>
<td>Logistic Regression</td>
<td>96.44%</td>
<td>96.95%</td>
<td>62.85%</td>
<td>76.26%</td>
<td>81.33%</td>
</tr>
<tr>
<td></td>
<td>Deep Learning</td>
<td>96.82%</td>
<td>97.07%</td>
<td>67.00%</td>
<td>79.27%</td>
<td>94.22%</td>
</tr>
<tr>
<td>Health Facts</td>
<td>Random Forest</td>
<td>98.69%</td>
<td>95.34%</td>
<td>85.7%</td>
<td>90.30%</td>
<td>95.11%</td>
</tr>
<tr>
<td></td>
<td>Decision Tree</td>
<td>97.38%</td>
<td>87.47%</td>
<td>83.10%</td>
<td>85.23%</td>
<td>93.38%</td>
</tr>
<tr>
<td></td>
<td>Logistic Regression</td>
<td>95.73%</td>
<td>96.17%</td>
<td>55.20%</td>
<td>70.14%</td>
<td>74.32%</td>
</tr>
<tr>
<td></td>
<td>Deep Learning</td>
<td>97.93%</td>
<td>99.23%</td>
<td>77.80%</td>
<td>87.22%</td>
<td>95.41%</td>
</tr>
</tbody>
</table>

Results demonstrated our models are capable of classifying opioid poisoning very well. We can achieve an F1 score of 79.27% for SPARCS (by deep learning) and 90.30% for Health Facts (by random forest). For precision, the best performance we can achieve is through deep learning: 97.07% for SPARCS and 99.23% for Health Facts, indicating the ability of models to identify only opioid poisoned patients. For recall, we can achieve 71.68% for SPARCS and 85.70% for Health Facts respectively for the best case, indicating the ability to find all the opioid poisoned patients in the dataset. The best AUC score we can achieve is 94.94% for SPARCS and 95.41% for Health Facts respectively.

For the SPARCS dataset, which comes with a smaller number of features, deep learning generates the best F1 score and a comparable AUC score with random forest. For Health Facts, which has more comprehensive features, random forest has the best F1 score, and deep learning generates the best AUC score.

**Feature Analysis**

It is vital to understand the importance of different features in the models, to support researchers or clinicians to exploit potential causes or trajectories of diseases. Based on our best traditional method random forest, we generated top most important features for both datasets in Table 3 and 4 respectively. The rank and feature category are also shown in the tables. Random forest calculates the importance of each feature by using Gini importance\(^1\). It is defined as the total decrease in node Gini impurity index weighted by the probability of reaching that node averaged over all decision trees of the ensemble. Gini impurity index is defined as:

\[
G = \sum_{i=1}^{n_c} n_c p_i (1 - p_i)
\]

where \(n_c\) is the number of classes, \(p_i\) is the ratio of this class, and \(G\) is the Gini impurity index. Random forest is an ensemble of decision trees. For every input sample, each decision tree will decide the path from root to leaf node for it, and the leaf will tell if the case is positive or negative. Then the model will ensemble the results of all decision trees to decide the final result.

In every node of the decision tree, there is a feature to decide the path after the node. So for each node, samples left for each class are different, and the Gini impurity index will be different. Then we have Gini importance as:

\[
I = G_{\text{parent}} - G_{\text{split1}} - G_{\text{split2}}
\]

\(G_{\text{parent}}\) is the Gini impurity index of the parent node, \(G_{\text{split1}}\) and \(G_{\text{split2}}\) are the Gini impurity indexes for two child nodes. Finally, the weighted summation of Gini importance of all nodes for one feature in all decision trees will be the final importance of that feature.

For New York State SPARCS data, most important features are diagnosis codes. The top feature is “Sedative, hypnotic or anxiolytic dependence, continuous” (ICD-9 code 304.11, approximately mapped to ICD-10 code F13.20 “Sedative,
hypnotic or anxiolytic dependence, uncomplicated”). These disorders result from the non-medical use of medications known as sedatives, hypnotics, and anxiolytics. The second top feature is relevant to Hepatitis C (ICD-9 code 070.70 or ICD-10 code B19.20), which is commonly transmitted by needles shared among those who inject drugs non-medically34. The third top feature “other alteration of consciousness”. Common underlying causes of decreased consciousness includes drugs, alcohol, substance abuse, certain medical conditions, epilepsy, low blood sugar, etc. Note that marijuana misuse (“cannabis abuse, unspecified”) ranks top 4. There are multiple top features relevant to chronic pain (“other chronic pain”, “chronic pain syndrome”, “lumbago”). Injury from car accidents or third-degree burn are also among the top 20 features.

Table 3. Top 20 Important Features for Prediction on SPARCS Dataset.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Sedative, hypnotic or anxiolytic dependence, continuous</td>
<td>1</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Unspecified viral hepatitis C without hepatic coma</td>
<td>2</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Other alteration of consciousness</td>
<td>3</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Cannabis abuse, unspecified</td>
<td>4</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Anxiety state, unspecified</td>
<td>5</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Unspecified drug dependence, unspecified</td>
<td>6</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Sedative, hypnotic or anxiolytic abuse, unspecified</td>
<td>7</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Toxic effect of ethyl alcohol</td>
<td>8</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Other chronic pain</td>
<td>9</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Pedestrian injured in collision with two- or three-wheeled motor vehicle</td>
<td>10</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Altered mental status</td>
<td>11</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Other, mixed, or unspecified drug abuse, unspecified</td>
<td>12</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Chronic pain syndrome</td>
<td>13</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Acute respiratory failure</td>
<td>14</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Lumbago</td>
<td>15</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Poisoning by other specified central nervous system stimulants</td>
<td>16</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Depressive disorder, not elsewhere classified</td>
<td>17</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Poisoning by benzodiazepine-based tranquilizers</td>
<td>18</td>
</tr>
<tr>
<td>Procedure</td>
<td>Emergency department visit for the evaluation and management of a patient</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 4. Top 20 Top Predicting Features in Health Facts Dataset.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Event</td>
<td>Pulse</td>
<td>1</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Heart Rate</td>
<td>2</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Advanced Directive</td>
<td>3</td>
</tr>
<tr>
<td>Procedure</td>
<td>Under New or Established Patient Emergency Department Services.</td>
<td>4</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Pulse Posterior Tibial Right</td>
<td>5</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Tobacco Use (Number of Years)</td>
<td>6</td>
</tr>
<tr>
<td>Medication</td>
<td>Sodium Chloride</td>
<td>7</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Respiratory Rate</td>
<td>8</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Other, mixed, or unspecified drug abuse, unspecified</td>
<td>9</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Triage Notes</td>
<td>10</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Essential (primary) hypertension</td>
<td>11</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Pain Scale Score</td>
<td>12</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Temperature (Route Not Specified)</td>
<td>13</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Fall, History of</td>
<td>14</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Pain Symptom Date of Onset</td>
<td>15</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>BMI, Body Mass Index</td>
<td>16</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Pulse Dorsalis Pedis Left</td>
<td>17</td>
</tr>
<tr>
<td>Procedure</td>
<td>Emergency department visit for the evaluation and management of a patient</td>
<td>18</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Tobacco use disorder</td>
<td>19</td>
</tr>
<tr>
<td>Clinical Event</td>
<td>Acetaminophen</td>
<td>20</td>
</tr>
</tbody>
</table>
The Health Facts database has much richer features compared to SPARCS data. It includes many clinical events and medications in addition to diagnoses and procedures. Interestingly, basic measurements such as heart rate (Pulse, Heart Rate, Pulse Posterior Tibial Right, or Pulse Dorsalis Pedis Left) and respiratory rate are among top predicting features, as taking too much of opioid can lead to a slow heart rate and difficulty maintaining breathing. Temperature, pain scale, history of fall, and BMI are also among the top 20 features.

Table 5 summarizes the aggregated importance of each category of features on contributing to the prediction. For SPARCS, diagnosis dominates the contribution, and the procedure follows. Demographic information also contributes to the prediction. For Health Facts, the descending order of importance of the predictor categories is clinical events, diagnosis, medications, procedure and demographic information.

Table 5. Importance for each feature category.

<table>
<thead>
<tr>
<th>SPARCS</th>
<th>Health Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Procedure</td>
</tr>
<tr>
<td>78.07%</td>
<td>15.27%</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Procedure</td>
</tr>
<tr>
<td>25.14%</td>
<td>6.17%</td>
</tr>
</tbody>
</table>

Experiment on Using Primary Data only for the Prediction

Since diagnosis codes are secondary data for billing purposes, we were interested in studying the performance with only primary data. We took the Health Facts dataset and removed all diagnosis codes and then built the same models. Table 6 shows the result before and after removing diagnosis features. By removing the diagnosis data, we could still achieve a good performance on prediction with similar results. Random Forest achieves slightly higher AUC without diagnosis codes.

Table 6. Performance of models before and after removing diagnosis features on Health Facts.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Have Diagnosis Codes</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>Yes</td>
<td>98.69%</td>
<td>95.34%</td>
<td>85.70%</td>
<td>90.30%</td>
<td>95.11%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>97.41%</td>
<td>96.25%</td>
<td>74.40%</td>
<td>85.56%</td>
<td>95.29%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>Yes</td>
<td>97.38%</td>
<td>87.47%</td>
<td>83.10%</td>
<td>85.23%</td>
<td>93.38%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>96.98%</td>
<td>93.15%</td>
<td>72.10%</td>
<td>81.29%</td>
<td>91.63%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Yes</td>
<td>95.73%</td>
<td>96.17%</td>
<td>55.20%</td>
<td>70.14%</td>
<td>74.32%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>95.73%</td>
<td>96.66%</td>
<td>55.00%</td>
<td>70.10%</td>
<td>75.69%</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>Yes</td>
<td>97.93%</td>
<td>99.23%</td>
<td>77.80%</td>
<td>87.22%</td>
<td>95.41%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>97.35%</td>
<td>98.36%</td>
<td>72.00%</td>
<td>83.14%</td>
<td>94.97%</td>
</tr>
</tbody>
</table>

Discussion

Data and evidence based studies hold high potential for studying the problem of opioid epidemic in the U.S. With the wide availability of Electronic Health Records, predictive modeling provides a powerful approach to automatically predict the risks of opioid overdoses. Compared to previous related work, for example Che et al. try to classify opioid overdose patients into short and long term users, our work is to identify potential opioid overdose patients among all patients. To our best knowledge, this is the first work on machine learning based opioid overdose prediction with large scale EHR data. Instead of limiting the method to a specific model, we proposed multiple models and compared their performance.
While claims data such as SPARCS has limited information, our experiments demonstrate that they can still achieve reasonable prediction results. More comprehensive EHR data such as Health Facts has richer information and thus can achieve very promising results. Our methods can achieve best recall with the random forest method (precision: 95.3%, recall: 85.7%, F1 score: 90.3% and AUC: 95.11%), and best precision with deep learning (precision: 99.2%, recall: 77.8%, F1 score: 87.2% and AUC: 95.41%).

The analysis of features also provides interesting findings. For example, in SPARCS-based claims data, Hepatitis C is an indicator of potential non-medical opioid use, as Hepatitis C infection is frequently infected due to the sharing of needles and syringes for drug use. Marijuana misuse, tobacco use and alcohol are also linked to opioid overdose. Chronic pain management and treatment of acute pain can lead to opioid overdose. With more comprehensive EHR data, we find many clinical events such as measurements of heart beat, respiratory rate and pain scale that can help the prediction, and the clinical events dominates the contribution for prediction. By removing secondary data such as diagnosis codes, we find that our model can still achieve comparable performance. Our study also demonstrates that demographic information and sources of payments are helpful factors for the prediction.

One important information source that could enhance our models is clinical notes, which are missing from datasets in this study. In our future work, we will include clinical notes by applying natural language processing technologies. Another limitation of our method is that we did not employ any specific imputation method to deal with the missing value problem. To include more features with many missing values, we will employ different imputation methods. Besides, for convenience in this work, we took all patients that has not been diagnosed as opioid overdose as negative patients, but actually we don’t know about their future, we will try to apply more reasonable way to identify positive and negative patients. For the deep learning models, we currently have a relatively simpler architecture which does not take advantage of the temporal factors of the data and visits. We will explore RNN method and combine both outpatient and inpatient visits for future studies.

Conclusion

The opioid epidemic has become a national emergency for public health in the United States. Identifying patients of high risk of opioid overdose can provide smarter and safer clinical decisions for treatment, and potentially improve current clinical practices for pain management. Our studies on machine learning based predictive models for opioid overdose prediction show promising results with both claims data and comprehensive EHR data. It demonstrates that an AI based approach, if implemented at clinical side, can achieve automated prediction with high accuracy, and provide an opportunity for AI assisted prediction to support healthcare providers to combat opioid crisis.

References

32. Chollet F. Keras. 2015.
Longitudinal analysis of social and behavioral determinants of health in the EHR: exploring the impact of patient trajectories and documentation practices

Daniel J. Feller, MA¹, Jason Zucker, MD², Oliver Bear Don’t Walk, MS¹, Michael T Yin², MD, Peter Gordon, MD³, Noémie Elhadad, PhD¹

¹Department of Biomedical Informatics, Columbia University, New York, NY, USA; ²Division of Infectious Diseases, Department of Medicine, Columbia University, New York, NY USA

Abstract

Social and behavioral determinants of health (SBDH) are environmental and behavioral factors that impede disease self-management and can exacerbate clinical conditions. While recent research in the informatics community has focused on building systems that can automatically infer SBDH from the patient record, it is unclear how such determinants change over time. This study analyzes the longitudinal characteristics of 4 common SBDH as expressed in the patient record and compares the rates of change among distinct SBDH. In addition, manual review of patient notes was undertaken to establish whether changes in patient SBDH status reflected legitimate changes in patient status or rather potential data quality issues. Our findings suggest that a patient’s SBDH status is liable to change over time and that some changes reflect poor social history taking by clinicians.

Introduction

Social and behavioral determinants of health (SBDH) are environmental and behavioral factors such as unstable housing and substance use disorders that often impede disease self-management and can lead to or exacerbate existing comorbid conditions. Due to the established impact of SBDH on health outcomes for persons living with chronic disease, health systems are increasingly attuned to these determinants and the clinically meaningful information they provide. However, evidence suggests that providers often struggle to retrieve information related to SBDH, and that those providers who are unaware of SBDH provide worse quality care.

Integrating SBDH data into electronic health records (EHRs) has generated widespread interest among health systems and healthcare providers. The informatics community has focused on three approaches to integrating SBDH into EHRs: expanding data collection to include structured data elements representing SBDH, associating community-level SBDH information with individual patients, and using information extraction and Natural Language Processing (NLP) methods to identify patient-level SBDH information contained in clinical documentation. SBDH such as smoking status, substance abuse, and homelessness have been the focus of recent NLP approaches, looking for identifying these at the encounter level, but have to date achieved modest-yet promising performance. However, there has been little work so far on characterizing a patient’s SBDH status beyond a single encounter.

It is unlikely that a patient’s documented SBDH status is invariably consistent with the patient’s true state. Patients are often hesitant to disclose sensitive information such as sexual orientation and substance use to healthcare providers, and may be less likely to share sensitive information with non-physician providers. In addition, the quality of social history taking by clinicians is variable, and providers are liable to make incorrect assumptions about their patient’s health. As a result, a patient’s SBDH status recorded in the patient record may reflect inaccuracies attributable to phenomena inherent to clinical documentation of sensitive information.

There is little knowledge of how social and behavioral determinants of health as expressed in patient records change through time. While it has been established that an individual’s SBDH may change across the life course, it is unclear whether the information in EHRs accurately represent such changes. To our knowledge, the only relevant study that examined changes in SBDH documentation was conducted by Bejan et. al. in 2017, which observed cyclic transitions between the at-risk and homeless categories among homeless patients, and less frequent transitions among individuals with stable housing. In this paper, we conduct a longitudinal analysis of multiple SBDH as expressed in the patient record and attempt to answer several important questions. First, we assess whether SBDH change through time and
estimate their respective rates of change. Second, we conduct a survival analysis to examine the timescale of these changes. Third, we employ both quantitative and qualitative methods to examine potential data qualities issues related to SBDH.

**Dataset**

In this section, we describe the methods used to curate and analyze a longitudinal gold-standard corpus of clinical notes. A protocol for manual annotation of clinical notes was developed and followed by multiple rounds of annotation. The study described herein was approved by the Institutional Review Board at CUIMC and patient informed consent was waived due to the retrospective nature of the study.

**Development of Annotation Guidelines for SBDH**

Two physicians experienced in the prevention and treatment of HIV reviewed the biomedical literature to identify social and behavioral determinants that increased an individual’s risk for acquiring HIV. 29 individuals risk factors were selected and included sexual orientation, housing status, alcohol use and drug use which are the focus of the study described herein.

We obtained document-level annotations for the aforementioned SBDH domains and risk factors and thus considered all SBDH as binary variables. Annotators were instructed to review the entire length of each clinical document and label the presence of both the high-level domain and granular SBDH label. For example, the phrase “patient denies alcohol use” would be result in a positive label for the high-level SBDH domain ‘alcohol use documented’ but a negative label for ‘active alcohol use’. The annotation guide can be downloaded at [https://github.com/danieljfeller/SBDSH](https://github.com/danieljfeller/SBDSH).

**Creation of Gold-Standard Corpus for Longitudinal Analysis**

A corpus of clinical notes was obtained from the clinical data warehouse at Columbia University Irving Medical Center (CUIMC), a large academic medical center in New York City. For this study, we obtained all individual notes associated with 32 randomly sampled individuals with HIV who received care at CUIMC. The study cohort was characterized by analyzing structured demographic data extracted from the EHR system at CUIMC. Additional details on this cohort are described elsewhere.

Annotation was conducted by 1 physician and 2 medical students who were instructed to manually review every clinical document included in the patient record. An initial set of 3 longitudinal patient records were collaboratively coded by all annotators for training purposes; 76 notes were annotated by each of the 3 reviews and a Kappa of 0.736 was observed across all SBDH labels. Throughout the annotation process, annotators relied on the guidelines described above.

**Creation of Gold-Standard Corpus for Data Quality Analysis**

We collected a larger gold-standard corpus for analyzing data quality of SBDH documentation. In contrast to the longitudinal corpus described above, we included in this cohort HIV+ individuals associated with only a portion of their clinical notes annotated and required only that individuals had 2 or more notes with SBDH documentation. These notes were annotated for the purposes of a previous study.

**Methods**

The analysis described herein focused on 4 SBDH; sexual orientation, housing status, alcohol use and drug use. Encounters with confirmed documentation of SBDH were isolated and analyzed to examine changes in a patient’s SBDH status. and potential data quality issues. We then manually reviewed pairs of notes authored on the same-day with conflicting documentation to identify possible sources of data quality issues related to SBDH.

**SBDH Preprocessing**

A distinct dataset was created for each SBDH of interest and included only notes where the respective high-level SBDH was documented. For example, to be included in the analysis of ‘drug use’ status, it was necessary that a note discuss drug use (even if drug use was a negative label, as in “no history of substance use”). This way, all notes in that SBDH dataset had explicit discussion of that SBDH and either positive or negative findings for that SBDH. Notes that did not discuss the SBDH were not included; absence of a certain SBDH in a clinical note most often reflects the fact that this domain was not discussed by patient and provider, rather than evidence that the patient is a negative case.
Survival Analysis

To assess the rate of change in SBDH status, we simply parameterized each annotated document using sequence time (e.g., visit\_time\_1, visit\_time\_2, etc. where visit\_time\_i represents the time between the first and i\textsuperscript{th} visit in the longitudinal record of a patient) and estimated the likelihood that a patient at visit\_i would transition to a different state (e.g., ‘never used alcohol’ to ‘actively using alcohol’), as documented within the documentation at the next visit. This analysis was conducted using the corpus generated by a comprehensive annotation of the entire longitudinal record of the 32 HIV+ individuals in our cohort.

We used survival analysis to analyze the expected duration of time associated with changes in patient SBDH status. Observation periods were established between adjacent notes in a patient’s longitudinal history and time was measured in days. An event was defined as any change in SBDH status observed in the subsequent note. Observations were (right) censored when the subsequent note was observed with the same SBDH status as observed in the preceding (index) note. A survival function was estimated for each SBDH using the Kaplan-Meier estimate and can be interpreted as the fraction of clinical notes observed at time \( T \) with an observed change in SBDH status documented in the subsequent note:

\[
\hat{S}(t) = \prod_{t_i \leq t} \left( 1 - \frac{d_i}{n_i} \right)
\]

where \( d_i \) is the number of notes with subsequent SBDH changes at time \( t_i \) and \( n_i \) is the number notes not associated with any subsequent SBDH change (and who have not yet been censored) at time \( t_i \).

Quantitative Analysis of Data Quality

In order to assess data quality in SBDH documentation, we used a larger cohort of patients who were required only to have multiple notes manually annotated. Using the parameterization of sequence time described above, we considered illegitimate transitions to be those that were chronologically impossible; for example, a patient could be documented as having ‘never used alcohol’ subsequent to being documented as ‘actively using alcohol’.

In addition, we identified same-day conflicts in SBDH documentation by isolating clinical notes that were written on the same day by distinct providers. Similar to previous analyses, we required that all notes under consideration have confirmed documentation of the relevant high-level SBDH domain. Same-day conflicts were defined as observed discrepancies in SBDH status (e.g., documentation of “no active alcohol use” and “active alcohol use”).

Qualitative Analysis of Data Quality

We manually reviewed 20 note pairs that exhibited same-day conflicts in patient SBDH status with the goal of developing an understanding of the sources of data quality problems. 5 note pairs were gathered from each of the four SBDH analyzed herein. The sources of data quality issues were identified using a set of annotation guidelines we created for this analysis. Data quality issues reflected 1) inappropriate use of copy & paste if content was duplicated across notes of the same type, 2) an inaccurate problem list if outdated information contained in structured clinical data was automatically inserted into narrative free-text, 3) variable history taking wherein the note with positive documentation contained significantly more information regarding SBDH compared to the note with negative documentation (e.g. provider documents a minimal social history and does not inquire about specific substances, for example”), 4) a patient hesitant to disclose sensitive information when it was clear that both notes contained a detailed social history but that the patient gave conflicting answers across the notes, and 5) the use of a standard note template which automatically inserted negative and formulaic documentation of SBDH status.

Results

Cohort Characteristics

3273 clinical documents associated with 32 HIV+ individuals were manually annotated and included in the longitudinal corpus. All available clinical notes associated with these patients was annotated with a range of 11 to 473 notes per patients (median 50, mean 102). The longitudinal histories of patients in this analysis were of variable lengths and ranged from 196 days to 3146 days; the mean longitudinal history was 997 days and the median was 772 days.
75% of individuals were male (24) and the average age was 46 years with a standard deviation of 13.5 years. Race and ethnicity information was missing for a majority of the cohort, but 8 patients were documented as African American, 7 as Caucasian Hispanic, 2 as Caucasian non-Hispanic, and 1 as Asian.

The larger cohort used in the data quality analysis included 366 individuals with multiple annotated notes (4294 notes total). All available clinical notes associated with these patients was annotated with a range of 2 to 473 notes per patients (median 2, mean 11). The longitudinal histories of patients in this analysis were of variable lengths and ranged from 0 days to 3146 days; the mean longitudinal history was 469 days and the median was 174 days. 60.3% of individuals were male (221) and the average age was 53 years with a standard deviation of 12.2 years. Race and ethnicity information was missing for a majority of the cohort, but 71 patients were documented as African American, 58 as Caucasian Hispanic, 30 as Caucasian non-Hispanic, and 1 as Asian.

### Longitudinal Changes in SBDH

A state diagram illustrating changes of documentation status for alcohol use computed across the cohort of 366 individuals is presented in Figure 1, across 1077 pairs of consecutive notes.

Many patients in the longitudinal cohort of 32 individuals were observed to have temporal changes among all SBDH analyzed. Alcohol status was most likely to change across subsequent notes (23.3%), followed by drug use (10.4%), drug use (8.7%), and sexual orientation (1.1%). A chi-square goodness of fit test found a significant difference in these proportions ($\chi^2: 29.2, p < 0.001$).

A Kaplan-Meier plot that presents the 365-day survival function of the 4 SBDH analyzed in this study in presented in Figure 2. The y-axis represents the proportion of notes without changes in SBDH documentation, and the x-axis represents the number of days between each subsequent note. 365-day transition rates for housing status, drug use, and alcohol use were 39.6%, 30.6%, and 36.3%, respectively. The 365-day transition rate for sexual orientation was 6.9%.

### Data Quality in SBDH Documentation

While most of the transitions in SBDH status looked sensible, we observed illegitimate transitions in SBDH status as shown in Figure 1. Among 353 notes documenting patient SBDH status as ‘active alcohol use, 43 were followed by documentation of ‘never used alcohol’ (12.2%). Moreover, among 165 notes documented that the patient had ‘historical alcohol use’, 26 were followed by documentation of ‘never used alcohol’ (15.7%).

We also observed same-day conflicts in patient SBDH status. For instance, 23.2% of 56 same-day note pairs with alcohol status documented had conflicting indications of alcohol use, 21.2% of 52 same-day note pairs had conflicting indications of substance use, 6.8% of 44 same-day note pairs had conflicting indications of patient housing status, and 8.0% of 25 same-day note pairs had conflicting of sexual orientation. We also observed conflicts in SBDH documentation within a 7-day period. 12.6% of annotated notes associated with the same patient had conflicting
indications of alcohol use (N = 230). In addition, 14.3% of notes describing drug use exhibited the same changes (N = 265), as did 7.6% of notes describing housing status (N=198). 1.6% of notes with sexual orientation documentation exhibited same-week conflicts (N = 127).

Manual review of 20 same-day conflicts in SBDH revealed multiple sources of poor data quality with incomplete social history taking by clinicians being most common. Among the 20 manually reviewed note pairs with same-day conflicts in SBDH, 14 (70%) conflicts reflected variable quality of social history taking by clinicians, wherein the note with positive documentation contained significantly more detailed information regarding SBDH compared to the note with negative documentation. The majority of such instances entailed particularly brief social histories that may have reflected inaccurate assumptions and/or limited social history taking by healthcare providers. 2 note pairs exhibited evidence of an inaccurate problem list, wherein note content automatically generated from the patient’s EHR problem list conflicted with information contained in narrative free-text. 2 note pairs exhibited evidence that a patient was hesitant to disclose SBDH, as it was clear that both notes contained a detailed social history but that the patient gave conflicting answers across the notes. 1 note pair conflict reflected the use of a standard note template, which automatically declared negative SBDH status.

Discussion

The findings of this study suggest that social and behavioral determinant of health as expressed in the patient record exhibit changes over time. Our longitudinal analysis of 4 distinct risk-factors suggests that a patient’s SBDH status should be treated as a shifting, mutable variable in electronic systems. We also provide evidence that some changes in SBDH documentation reflect data quality issues and not actual changes in the patient state.

We present the first comprehensive longitudinal analysis of multiple SBDH as expressed in patient records. Four distinct SBDH were examined throughout the course of 32 patient records and exhibited varying rates of change. A patient’s recorded alcohol status was most likely to change, as 23.3% of adjacent encounters with documentation of alcohol use contained conflicting information. This may reflect the high prevalence of heavy episodic drinking among persons living with HIV, which has resulted in calls for repeated assessments of alcohol consumption in this population.\textsuperscript{30,31} While documentation of substance abuse and housing status were less likely to change compared to alcohol use, these SBDH exhibited changes across as much as 10% of adjacent encounters. Epidemiological studies suggest that these SBDH are likely to change; most individuals who are considered with unstable housing experience only transient periods of homelessness\textsuperscript{37}, and many HIV+ persons with substance abuse disorder engage in episodic
rather than sustained use.\textsuperscript{32,33} These findings suggest that automated approaches to inferring SBDH should not treat these variables as fixed patient characteristics and thus should reevaluate a patient’s SBDH status on a regular basis.

In contrast to alcohol use, substance abuse, and housing status, a patient’s recorded sexual orientation was unlikely to change, as 1.1% of adjacent encounters with documentation of sexual orientation contained conflicting information. This rate reflects the infrequent changes in sexual orientation observed among persons experiencing stigma and discrimination.\textsuperscript{34} This findings suggest that distinct SBDH are likely to change at different rates and thus may be reevaluated on different time scales.

Multiple findings indicated that some changes in patient SBDH status reflected data quality issues and not legitimate changes in the patient state. We observed a high frequency of implausible longitudinal changes in patient SBDH status, wherein a patient transitioned from an active status (e.g. active alcohol use) to having no history of active status (e.g. no history of alcohol use). In addition, we observed same-day conflicts in patient’s documented SBDH status. Our manual review of these discrepancies observed that most conflicts reflected the variable nature of social history taking. It has been established that some healthcare providers are reluctant to discuss sensitive issues with their patients, thereby limiting their ability to take a comprehensive social history.\textsuperscript{35} The implication of these data quality issues is that any attempt to characterize a patient’s SBDH status should not merely reflect the most recent documentation. Decision support systems that aggregate multiple instances of SBDH documentation may provide a more faithful representation of a patient’s SBDH status compared to data collected during a single encounter. Information retrieval and classification methods should utilize observation windows that leverage only recent EHR data\textsuperscript{36,37}, or weight decay techniques that model the decreasing relevance of data elements to computational phenotypes.\textsuperscript{38,39}

Future research should conduct a more comprehensive analysis by annotating all notes associated with a large corpus and use techniques such as mutual information to assess how the predictability of future SBDH status relative to existing documentation changes with time.\textsuperscript{40} Such an analysis would open the possibility for techniques that could accurately model the relevance of social and behavioral determinants of health documentation.\textsuperscript{41,42} In addition, more research is needed to better estimate the likelihood that changes in SBDH documentation reflect true changes in the patient state versus documentation errors. This could be accomplished by obtaining additional sources of data such as validated psychometric instruments or laboratory tests indicative of substance abuse.

Limitations

First, our methods did not enable us to quantify the proportion of SBDH changes that reflected true changes in the patient state and the proportion that reflected documentation errors. A key limitation of this study is our inability to estimate whether a patient’s SBDH status is accurate and note that obtaining this information is likely not possible using clinical notes alone. Second, while our annotators achieved a relatively high inter-rater reliability, there were likely some erroneous annotations and thus some temporal changes in SBDH status may reflect annotation errors rather than changes in documented SBDH status. Third, our findings were generated by analyzing data from a specific patient cohort treated at a single institution. The high prevalence of SBDH within the study cohort may have resulted in a higher frequency of SBDH changes.

Conclusion

Social and behavioral determinants of health as documented in the patient record are subject to change through time. Future approaches to automated inference of SBDH from EHRs should consider the fluid nature of these variables and the impact of data quality on their methods and results.

References


42. Dasu, T. & Weiss, G. Mining Data Streams. in Encyclopedia of Data Warehousing and Mining (2009).
Hypersphere clustering to characterize healthcare providers using prescriptions and procedures from Medicare claims data

Nathanael Fillmore, Ph.D.\textsuperscript{1,2}, Sergey D. Goryachev\textsuperscript{1}, and Jeremy C. Weiss, M.D. Ph.D.\textsuperscript{3}
\textsuperscript{1}Department of Veterans Affairs, Boston, Massachusetts, USA; \textsuperscript{2}Harvard Medical School, Boston, Massachusetts, USA; \textsuperscript{3}Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Abstract

We consider the task of producing a useful clustering of healthcare providers from their clinical action signature—their drug, procedure, and billing codes. Because high-dimensional sparse count vectors are challenging to cluster, we develop a novel autoencoder framework to address this task. Our solution creates a low-dimensional embedded representation of the high-dimensional space that preserves angular relationships and assigns examples to clusters while optimizing the quality of this clustering. Our method is able to find a better clustering than under a two-step alternative, e.g., projected K means/medoids, where a representation is learned and then clustering is applied to the representation. We demonstrate our method’s characteristics through quantitative and qualitative analysis of real and simulated data, including in several real-world healthcare case studies. Finally, we develop a tool to enhance exploratory analysis of providers based on their clinical behaviors.

Introduction

Exploratory analysis and visualization of health data are central to the advancement of clinical care. Clinical data are naturally high-dimensional and sparse because most medical concepts are not applicable to most individuals. For example, any individual physician performs a small subset of procedures, and any patient fills prescriptions for a small subset of medications. Our work seeks to provide effective representations of these settings by producing useful clusterings of high-dimensional sparse count data. Our focus is in the analysis of large-scale healthcare data arising from claims and health records databases, where we are interested in clustering patients or providers based on the number of time each billing code, procedure code, or drug prescription is observed.

Similarity between sparse, high-dimensional points is typically better measured through angular distances like cosine similarity rather than euclidean distance. As a result, many standard clustering methods will perform poorly in these tasks, and clustering algorithms in lower-dimensional representations have been developed to address this problem. For example, one popular approach adopts a two-step process: (a) learn a low-dimensional representation, (b) apply clustering over the low-dimensional representation, which is illustrated by the exemplar of principal components analysis (PCA) to find a low-dimensional representation followed by k-means to cluster points based on this representation\textsuperscript{1}. A variety of deep learning methods have also been used to find a low dimensional representation in the first step\textsuperscript{2}. For example, Tian et al.\textsuperscript{3} use stacked autoencoders to learn a representation, followed by k-means. Similarly, Trigeorgis et al.\textsuperscript{4} define a deep non-negative matrix factorization and use the representation thus learned to cluster data with k-means.

A major drawback of this two-step process is that the representation learned in the first step is not necessarily well suited to producing a clustering in the second step. For example, autoencoding is designed to minimize reconstruction loss, and a representation that minimizes reconstruction loss is not guaranteed – or even particularly likely – to map points to a meaningful clusterable space.

To address this issue, several methods have been defined that integrate both steps – reconstruction learning and clustering – in a single deep learning framework. This enables the representation to be informed by the clustering, not just by its ability to reconstruct data, and thus creates the possibility of finding both a more clusterable representation and a better clustering. For example, Xie et al.\textsuperscript{5} add a centroid-based soft clustering criterion to an autoencoder backbone and iteratively refine the autoencoder’s parameters so as to improve the quality of the clustering. Similarly, Yang et al.\textsuperscript{6} define a method that integrates an autoencoding deep neural network and a k-means clustering loss, and fits this model with an alternating stochastic gradient descent algorithm.
While these approaches show excellent performance on certain types of data, e.g. imaging, they are not well-suited to sparse count data. This is primarily due to their dependence on autoencoding with euclidean distances for representation learning. Even with additional constraints to encourage the representation to be well-clusterable, the autoencoders may not retain angular information relevant to a representation for high-dimensional sparse data. The use of centroid-based clustering compounds this problem because the identification of an appropriate centroid location may not be representative of any member of the cluster.

We argue that, instead of a centroid-based approach, an angular-based clustering is better suited for high-dimensional sparse data and that, therefore, an angular-based representation should be adopted. This representation constraint motivates our development of a novel clustering approach that embeds the representation on the surface of a hypersphere, where the angular distance from the axes or poles naturally provides cluster membership. While several support vector clustering algorithms have the option to incorporate angular distance into the kernel defining an embedded space\textsuperscript{7,8}, these methods do not preserve angular representation during clustering (instead the embedding is transformed into a graph), and these methods have not been shown to be scalable to the large sample sizes present in our applications.

**Our contribution.** We propose a method that learns a low-dimensional embedding with geometric constraints that enables simultaneous representation learning and cluster determination. We propose an architecture that simultaneously (1) optimizes signature reconstruction in an autoencoding framework, (2) embeds the signature on a low-dimensional hypersphere surface, and (3) assigns examples to clusters in a soft clustering by angular similarity. We further provide the following options. To account for the setting the case where we have expert knowledge about co-cluster membership, we provide a placeholder for an accessory loss to encode this information during optimization. The degree to which each component is prioritized can be adjusted by the user to enable enhanced exploratory analysis. To efficiently compare clusterings with different numbers of clusters, we also build our framework with a fixed hypersphere representation tied to lower-dimensional clustering hyperspheres.

We demonstrate our method’s characteristics and its advantages over previous work in simulations, and we illustrate the value of this approach through several case studies: provider characterization from Medicare claims and imaging characterization by radiology notes. We demonstrate the utility of our method with an informatics tool that characterizes providers based on their procedure and prescription signature in Medicare claims.

**Method**

We first describe the method and architecture with respect to inputs and outputs, then describe the distances and losses necessary for learning the representations, and finally specify customizations and implementation parameter choices.

Consider our input of a sparse count matrix $M$ with $N$ rows and $K$ columns. We are interested in clustering over rows (providers) and columns (prescriptions and procedures). We will consider clustering across rows and columns separately. First, we construct an autoencoder consisting of an encoder $\Psi_1$ and a decoder $\Psi_2$. For a row of matrix $M$, define the hidden representation $h$ of size $H$ such that $\Psi_1(x) = h$ and $M$ as the reconstruction given by $\hat{M} = [\Psi_2 \circ \Psi_1(x), \forall x \in M]$. Define $\Psi_3$ as the function transformation for the clustering module that takes as input $h$ and outputs cluster membership probabilities $c$, a non-negative vector of size $C$ that sums to 1: $\Psi_3(h) = c$. Figure 1 illustrates the framework for the neural network.

Next we define our losses which encode our desire to create a representation that maintains pairwise distances while endowing an angular distance for clustering. Unlike support vector clustering that uses all pairwise distances with complexity $O(N^2)$, our method uses pairwise batch distances $O(B^2)$ with batch size $B$, which alleviates the computational bottleneck when $N$ is large. Specifically, we propose to learn a hidden representation $h = \{h^c, h^||\}$ given by angle $h^c$ and norm $h^||$, such that the distance measure between examples in $M$ is approximately preserved in the angle $h^c$. To achieve this, we define batch pairwise similarity and its loss $L_{h^c, O}$ as a penalization to the autoencoder loss as follows. Define $h_B$ and $\hat{M}_B$ the representation and reconstructions for a batch of examples with sizes $B \times \{\cdot\}$ for $H$ and $M$ respectively. Let $\delta(x_i, x_j)$ be the pairwise distance measure.

Define $\delta_B$ the batch distance measure, i.e. $\delta_B(\{x_1, \ldots, x_B\}) = [\delta(x_i, x_j)]_{ij} \forall i, j \in \{1, \ldots, B\}$. Then, $L_{h^c, O} = \frac{1}{B} ||\delta_B(h^c_B) - \delta_B(\hat{M}_B)||_2^2$ captures the differences between the pairwise distances in $h^c_B$ and $\hat{M}_B$. 

409
Figure 1: Autoencoder with clustering module. Nodes are Input (I), Output (O), Hidden (H), and Cluster (C). H and O are constrained by a batch distance measure; we use cosine similarity for sparse count data. H \( h^* \) and C are constrained by batch cosine similarity that ties the angular representation to the cluster probability. See details in the text. The layers and connectivity are shown in the Table on the right. LReLU: leaky rectified linear unit

While \( L_{h^*O} \) encourages matching distances in \( h^* \) and \( M \), two vectors in \( h \) could have the same angle but different embedding locations, i.e., different \( h||\cdot|| \). Note this could be problematic because the hidden vectors are used to learn cluster membership probabilities \( c \), and different embedding locations mapping to the same cluster probability vector could induce unwanted within-cluster heterogeneity. To encourage approximate injectivity, we introduce the loss \( L_{|||\cdot|||} = 1 \) that penalizes hidden representations away from the surface of the unit norm hypersphere. We could enforce this as a hard constraint, however, the ability to violate the constraint may facilitate alignment of the embedded representation angle with that of the output space.

Now we leverage the angular embedding property of the hidden representation. The key is to note that for a probability vector that sums to 1, its element-wise square root vector has unit norm and can be interpreted as an angle. Therefore, like the previous loss, we can match the angular representation of \( c^\frac{1}{2} \) to that of \( h^* \) with batch cosine similarity.

Define \( c_B = \Psi_3(h_B) \) the batch of probability vectors indicating cluster membership. Note that the cosine similarities of rows of \( c_B^\frac{1}{2} \) are non-negative and we are not interested in incurring loss due to differences between 0 and negative cosine similarities. We define \( L_{c,h^*} = \frac{1}{B} ||\delta_B(c_B^\frac{1}{2}) - \max(\delta_B(h_B^*), 0)||^2_2 \).

We consider additional loss terms to assist customizable representation learning. First, we introduce an entropy loss term to encourage cluster probabilities to be spread across more than one cluster. This encourages large clusters with overlap to spread across multiple clusters to reveal subgroup characteristics and enables injecting belief about characteristics of the clustering. For optimization, it encourages exploration in cluster membership and could help avoid local optima.

Second and optionally, the framework can use accessory information to inform the clustering. Our primary framework encodes a single hidden representation for clustering, but does allow for specifying multiple numbers of clusters \( C \). In place of a single softmax, the terminal layer \( C \) can be a concatenation of possible numbers of clusters, e.g. \( \{2, 3, \ldots, 100\} \), where the softmax is applied to nodes corresponding to each of the possible numbers. Therefore, one can provide a cluster assignment of size \( C_{out} \) that informs the hidden representation and request a more or less granular clustering \( C_{out} \). Unlike hierarchical clustering, this extension produces a multiarchy.

Thus, our overall objective function is \( \sum_i \lambda_i L_i \) for \( L_i \in \{L_{M,M}, L_{c,h^*}, L_{h^*O}, L_{||h^*||=1}, L_{entropy}, L_{C_{out}}\} \). Unless otherwise specified, we set \( \lambda_i \) respectively: \( \lambda_i \in \{1, 1, 10^{-1}, 10^{-1}, 10^{-4}, 0\} \).

Finally, we describe our particular implementation choices, and suggest the reader to refer to the autoencoder frame-
Table 1: Quantitative evaluation on simulated data, relative to the true cluster labels. Data are generated as described in the text, with baseline parameters set at centroids=25, samples=1000, dims=1000, sod=0.01, explode=10000, and varied in each dataset as indicated. Data are clustered using our method (Sphere), and the closest neural embedding clustering techniques DCN, SAE+KM, and DEC. The best score is in boldface.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Adjusted Rand Index (ARI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sphere</td>
</tr>
<tr>
<td>baseline</td>
<td>1.00</td>
</tr>
<tr>
<td>samples=100</td>
<td>0.82</td>
</tr>
<tr>
<td>dims=100,000</td>
<td>0.64</td>
</tr>
<tr>
<td>explode=1,000,000</td>
<td>1.00</td>
</tr>
<tr>
<td>centroids=100</td>
<td>1.00</td>
</tr>
<tr>
<td>sd=0.1</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 2: Quantitative evaluation of variants of our method. Each column corresponds to a different variant of our own method, as defined in the text. The best score(s) is in boldface.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Adjusted Rand Index (ARI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sphere</td>
</tr>
<tr>
<td>baseline</td>
<td>1.00</td>
</tr>
<tr>
<td>samples=100</td>
<td>0.82</td>
</tr>
<tr>
<td>dims=100,000</td>
<td>0.64</td>
</tr>
<tr>
<td>explode=1,000,000</td>
<td>1.00</td>
</tr>
<tr>
<td>centroids=100</td>
<td>1.00</td>
</tr>
<tr>
<td>sd=0.1</td>
<td>0.25</td>
</tr>
</tbody>
</table>

work shown in Figure 1. In pre-processing we log transform each element using the function $f(x) = \log(1 + x)$ to prevent overemphasis of common events in the angular representation. We use cosine similarity for our distance measure between $h$ and $\hat{M}$ and between $h$ and $c$ as it is a natural distance representation for sparse count data. In the architecture, the permute-pool layer copies the tensor $P$ times, permutes the values, and performs max-pooling over the $P$ dimension with size 3 and stride 2. In our experiments we set $B = 128$, $P = 4$, and $H = 128$. We set $C$ to be twice the desired number of centroids, and in post-processing we iteratively merge the clusters identified based on maximum pairwise cluster distances.

Results

We evaluate our method on real and simulated data. Our results on simulated data show (1) that our method produces superior clusterings than other recently proposed deep clustering methods in this setting, and (2) how different components of our method affect the results of the method. Our case study evaluations on three real-world healthcare data show (1) the quality of our method’s clustering results, (2) the breadth of applicability of this method in a real-world context, and (3) the method’s scalability on large real-world data.

Quantitative evaluation

First, on simulated data generated to share properties of our target healthcare applications, we show that our method typically finds a clustering that, under several measures, is more similar to the ground truth labels from the simulation than competing methods are.

We generate data as follows. Centroids are sampled from a multivariate normal $\mathcal{N}(0, I)$ and normalized onto the unit sphere. We generate an equal number of samples for each centroid by adding noise distributed according to $\mathcal{N}(0, \sigma^2 I)$. Then we translate these points away from the origin by multiplying each point by an “explosion” factor $\kappa$ drawn from a random uniform on $[1, \kappa]$. The purpose of this step is to make the data more similar to that in our
healthcare applications, where patients and providers often have varying, i.e., non-constant, norm of total encounter counts.

Thus, the simulation is parameterized by the number of centroids, the number of dimensions, the explode factor, the number of samples, and the standard deviation of the noise. We set each parameter to a baseline level and vary one at a time to test our algorithm.

**Comparison to other methods.** We cluster the data using our hypersphere clustering method (“Sphere”), as well as three leading deep clustering frameworks that are similar to our method in combining autoencoders and clustering criteria: Stacked Autoencoder plus K-means (SAE+KM)\(^3\), Deep Embedded Clustering (DEC)\(^5\), and Deep Clustering Network (DCN)\(^6\). We evaluate results of these clustering methods using the Adjusted Rand Index (ARI) relative to the ground truth labels from the simulation.

Results are shown in Table 1. Our method produces better clusterings than the other methods on all six simulated datasets, as measured by ARI. Of note, the two-step SAE+KM method performs approximately as well as the integrated DCN and DEC methods on this data. This suggests that merely integrating clustering and representation learning into a single framework is not enough on its own to produce a good clustering; rather, it is also essential in this context for the representation to capture the relevant geometry of the data, as our method does.

**Comparison to variants of our method.** Next, we compare our method to variants of our method. These variants are created by omitting different components of the method one at a time, and thus, this evaluation demonstrates the importance of each component.

In the first variant (“Encoder”), we remove the decoder from the framework. Doing so results in worse clustering performance on all six datasets. The possible limitation of decoder removal is that it no longer needs to preserve the information in the input because the network does not need to reconstruct the input.

In the second variant (“HKM”), we run k-means in the hidden, embedded space our method constructs, instead of using our method’s clustering module. The embedded space is not designed to work well with k-means, and in fact, is constrained to have similar geometry to the original space, where k-means performs poorly. Likely for this reason, the HKM variant performs worse than our method on all datasets.

In the third variant (“Clu-t-SNE”), we substitute a t-SNE manifold constraint for \(L_{h=O}\), with the exception that to conform with the overall architecture of our method, instead of computing all \(O(N^2)\) pairwise distances, we define the KL-loss based on \(O(B^2)\) batch pairwise distances. This variant slightly outperforms our method when sample size is small (“samples=100”), but otherwise performs worse than our method.

The final variant (“Monotonic”) adds a constraint to our method rather than removing a component. Specifically, motivated by autoregressive flow literature\(^9\), we apply positive weight constraints and strictly monotonic activations (already present) to our network to get a autoregressive network. This variant outperforms ours in one case, but otherwise achieves equal or worse performance.

**Case study: Biclustering healthcare providers, prescriptions, and procedures with Medicare data**

The quantitative results in the previous section show that our method performs quite well on simulated data that we constructed to have properties of real-world high-dimensional sparse count healthcare data. In this section, we complement these quantitative results with a case study of provider characterization from procedure and prescriptions claims in Center for Medicare and Medicaid Services (CMS) Part D data. We compare our method’s results to those obtained by k-means, and use our method to explore the clusters identified.

**Objectives and Data.** The purpose of this case study is to cluster healthcare providers, prescriptions, and procedures based on data obtained from CMS claims, with the ultimate objective to gain insights on providers’ patterns of care based on this clustering, as well as on groupings of prescription and procedure use. To that end, we obtained Medicare Provider Utilization and Payment Data: Part D Prescriber Summary Table CY2015\(^10\), which tabulates all prescriptions and procedures given under the Medicare Part D program in 2015 in the United States. This data consists of approximately \(10^9\) providers, \(10^4\) procedures, and \(10^3\) drugs. We used our method and k-means to separately
Figure 2: Provider clustering by (a) hypersphere clustering, (b) k-means. Each column corresponds to a cluster and each row a specialty (only the most common ones are shown). We display the top three specialties associated with providers in that cluster, and order the axes according to number of providers and size of cluster. Relative to K-means, hypersphere clustering provides clusters with more diversity in top three specialties.

Comparison to k-means. The clusterings formed by our method and by k-means, each with 20 clusters, are shown in Figure 2, where the top three specialties within the identified clusters are visualized. The full list of included specialists per cluster is shown in the Appendix. On examination of these results, our method produces qualitatively better clusters than k-means does. For example, our clustering consistently includes a larger fraction of specialists in specialist clusters, e.g., Emergency Medicine (20k, Cluster 39), Dentist (85k, Cluster 25), and Psychiatry (21k, Cluster 38). Our clustering, unlike k-means, also identifies clean obstetrics and hematology oncology clusters. Although k-means identifies a cardiology and interventional cardiology cluster that our method does not initially identify, our method identified this cluster and merged it (Cardiology (18k), Nurse Prac (3k), Internal Medicine (2k)) into the internal medicine cluster (Cluster 38) in post-processing.

Our method also provides insight in regard to providers in specialties that are not as common. For example, our method’s urology cluster also includes a large fraction of the radiation oncologists in our dataset (see Appendix).
Figure 3: Left: histogram of nurse practitioner counts by cluster. Cluster names are given by clinical assessment of top 10 prescriptions. Right: t-SNE representation of 2000 randomly selected nurse practitioners. Cluster membership is indicated by color, and the axes are the dimensions of the t-SNE embedding.

detailed assessment of this cluster, we found that urology medications tamsulosin and finasteride were most commonly prescribed in this cluster and that radiation oncologists most commonly prescribed tamsulosin, followed by hydrocodone/acetaminophen and dexamethasone, possibly for prevention and treatment of complications of radiation therapy. Our clustering identified radiation oncologists and urologists as being similar according to the drugs they commonly prescribe, a finding that would not be identified through the use of a standard ontology alone.

To investigate our method’s biclustering (clustering on dimensions of provider and prescriptions), we investigated three questions: (1) does the clustering provide insight into how nurse practitioners provide care, (2) similarly, can we differentiate among family practice and internal medicine doctor subtypes, and (3) is the pain management cluster identifying commonality in opioid prescribing despite provider membership to many specialties?

Nurse practitioners. Nurse practitioners (NPs) tend to provide care in care units defined by medical specialties and subspecialties, but the CMS does not provide characteristics at that granularity. To characterize the type of care they provide, we can investigate their membership to identified clusters. Figure 3 (left) shows the breakdown of membership across clusters. We pulled the top 10 prescriptions administered by NP cluster members, as defined by mean log(1+·), and used our clinical expertise to characterize the care provided. Evidently, the NPs do cluster based on prescription behavior along lines of medical specialties. The top 10 medications for each cluster are provided in the Appendix. To empirically demonstrate the variation in the NP clustering, Figure 3 (right) shows a t-SNE(cos) plot of a random subset of 2000 NPs to illustrate their similarity (manifold distance) alongside their cluster membership (color). This demonstrates that the method does provide cluster separation and could be used to select NPs based on approximate specialization for future investigation.

Family practice and internal medicine. Similar to nurse practitioners, the specialty titles “family practice” and “internal medicine” are underdifferentiated. To investigate these providers’ prescription patterns, we inspected the FP/IM predominating clusters for differences. Figure 4(a) shows the breakdown of the providers across clusters, with labels corresponding to average care descriptions of the members in those groups. We selected three of the clusters that roughly corresponded to IM emphasis cardiology (IMC), IM emphasis infectious disease (IMID), and FP emphasis infectious disease (FPID). We compared the three clusters by listing the three prescriptions with the largest mean log(1+·) difference in prescribing.

- Compared to FPID providers, IMC providers prescribed more general IM medication (“levothyroxine”, “atorvastatin”, “lisinopril”) and fewer antibiotics (“levofloxacin”, “amoxicillin-clavulanate”, “azithromycin”).
- Compared to IMID providers, IMC providers prescribed more general internal medicine medication
Figure 4: Left: histogram of FP and IM provider counts by cluster. Cluster names given by clinical assessment of top 10 prescriptions. Right: snapshot of tool showing cluster membership for a selected physician, demo at http://4e46.net/docproc/.

{“atorvastatin”, “levothyroxine”, “lisinopril”} and fewer outpatient antibiotics {“clindamycin”, “amoxicillin”, “doxycycline”}.

- Compared to FPID, IMID providers prescribed stronger community-acquired pneumonia antibiotics {“levofloxacin”, “azithromycin”, “amoxicillin-clavulanate”} and fewer older and narrow-spectrum antibiotics and community care medications {“clindamycin”, “amoxicillin”, “ranitidine”}.

Pain management. Table 3 contains, by specialty, the top three medications prescribed by providers in the pain management cluster. The clustering demonstrates that, irrespective of the specialty, these providers prescribe large quantities of various opioids, including oxycodone, hydrocodone, morphine, fentanyl and others. We also compared the opioid-prescribing cluster against the other two clusters containing over 100 pain management specialty providers by listing the three prescriptions with the largest mean log(1 + ·) difference.

- Compared to the first cluster, the opioid-prescribing cluster prescribed more {“oxycodone”, “morphine”, “gabapentin”} and less {“celphalexin”, “prednisone”, “tramadol”}.

- Compared to the second cluster, the opioid-prescribing cluster prescribed more {“oxycodone”, “morphine”, “gabapentin”} and less {“ibuprofen”, “acetaminophen with codeine”, “azithromycin”}.

We developed a tool at http://4e46.net/docproc/ to facilitate learning about providers, their commonalities and memberships based on the their procedure and prescription signature in Medicare Part D claims data. This tool is illustrated in Figure 4 (right). An interested user can search by provider, cluster, or specialty to better understand group characteristics as well as the signature of practice for individual practitioners.

Other Case Studies

We also carried out several additional case studies. These case studies (1) demonstrate flexibility of the method with respect to number of desired clusters, (2) demonstrate the ability of our method to work well on real-world datasets across domains including text.

Single-representation multiple clustering. To illustrate single-representation multiple clustering, we run the simulation variant with 100 samples and 10 clusters, with samples 0-9 in cluster 0, 10-19 in cluster 1, and so on. Figure 5
Table 3: Top three medications by specialty for providers in the opioid-prescribing cluster.

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Prescriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anesthesiology</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Emergency medicine</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, prednisone</td>
</tr>
<tr>
<td>Family practice</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, morphine sulfate</td>
</tr>
<tr>
<td>General practice</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, gabapentin</td>
</tr>
<tr>
<td>General surgery</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Hematology/oncology</td>
<td>oxycodone hcl, morphine sulfate, ondansetron hcl</td>
</tr>
<tr>
<td>Hospice and palliative care</td>
<td>oxycodone hcl, morphine sulfate, fentanyl</td>
</tr>
<tr>
<td>Internal medicine</td>
<td>hydrocodone/acetaminophen, oxycodone hcl, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Interventional pain management</td>
<td>oxycodone hcl, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Medical oncology</td>
<td>oxycodone hcl, gabapentin, hydrocodone/acetaminophen</td>
</tr>
<tr>
<td>Neurology</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, gabapentin</td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, morphine sulfate</td>
</tr>
<tr>
<td>Nurse practitioner</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Obstetrics/gynecology</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Orthopedic surgery</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Otolaryngology</td>
<td>oxycodone hcl, fluticasone propionate, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Pain management</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Physical medicine and rehabilitation</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, gabapentin</td>
</tr>
<tr>
<td>Physician assistant</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Plastic and reconstructive surgery</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Podiatry</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Radiation oncology</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
<tr>
<td>Student</td>
<td>oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen</td>
</tr>
</tbody>
</table>

Cluster Procedure Count

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Procedure</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>CT ABDOMEN W/CONTRAST</td>
<td>6393</td>
</tr>
<tr>
<td>0</td>
<td>CT CHEST W/CONTRAST</td>
<td>5350</td>
</tr>
<tr>
<td>0</td>
<td>CT ABDOMEN W/O CONTRAST</td>
<td>3892</td>
</tr>
<tr>
<td>1</td>
<td>CHEST (PORTABLE AP)</td>
<td>13010</td>
</tr>
<tr>
<td>1</td>
<td>CHEST (PA &amp; LAT)</td>
<td>611</td>
</tr>
<tr>
<td>1</td>
<td>CHEST (SINGLE VIEW)</td>
<td>108</td>
</tr>
<tr>
<td>2</td>
<td>CHEST (PORTABLE AP)</td>
<td>11938</td>
</tr>
<tr>
<td>2</td>
<td>CHEST (PA &amp; LAT)</td>
<td>1693</td>
</tr>
<tr>
<td>2</td>
<td>CT CHEST W/O CONTRAST</td>
<td>11938</td>
</tr>
<tr>
<td>3</td>
<td>CHEST (PORTABLE AP)</td>
<td>14512</td>
</tr>
<tr>
<td>3</td>
<td>CHEST PORT. LINE PLACEMENT</td>
<td>1725</td>
</tr>
<tr>
<td>3</td>
<td>TRAUMA #3 (PORT CHEST ONLY)</td>
<td>217</td>
</tr>
<tr>
<td>4</td>
<td>BABYGRAM (CHEST ONLY)</td>
<td>3117</td>
</tr>
<tr>
<td>4</td>
<td>NEONATAL. HEAD PORTABLE</td>
<td>2782</td>
</tr>
<tr>
<td>4</td>
<td>P BABYGRAM (CHEST ONLY) PORT</td>
<td>1609</td>
</tr>
<tr>
<td>5</td>
<td>LIVER OR GALBLADDER US (SINGLE ORGAN)</td>
<td>4522</td>
</tr>
<tr>
<td>5</td>
<td>ABDOMEN U.S. (COMPLETE STUDY)</td>
<td>2102</td>
</tr>
<tr>
<td>6</td>
<td>CT C-SPINE W/O CONTRAST</td>
<td>4632</td>
</tr>
<tr>
<td>6</td>
<td>T-SPINE</td>
<td>1091</td>
</tr>
<tr>
<td>6</td>
<td>L-SPINE (AP &amp; LAT)</td>
<td>937</td>
</tr>
</tbody>
</table>

Figure 5: Left: multiple clustering outputs ($C = 2$ to $C = 20$) from a single representation, with y-axis indicating samples ordered by ground truth cluster membership each of size 10, and x-axis indicating the number of clusters found divided by the number available to the algorithm, and color indicating cluster membership within column. Right: Top three radiology descriptions per cluster (first seven) with procedure counts based on MIMIC radiology notes. Illustrates the predicted cluster assignments for different $C$ along the x axis, with colors for a given $C$ determining co-cluster membership and with colors for different $C$ learned to be close to the previous cluster color but far from other cluster colors. The figure demonstrates recovery of the ground truth clusters (10 of them) and also provides clusterings identified when $C \neq 10$, noting that the algorithm may elect to place 0 members in a cluster if the data warrants it.
Characterizing radiology notes. In our final case study, we sought to organize a large collection of radiology notes based on the notes’ contents, for individuals with critical care needs. We used de-identified radiology notes from MIMIC III v1.4, a natural language processing pipeline, and a bag of words representation to cluster radiology notes. The note descriptions provided are specific in some senses, e.g. anatomical: "X-ray of left foot 4th digit, two views", but nonspecific in other senses, e.g. by indication: “CT chest w/o contrast” for cardiac, pulmonary, or gastrointestinal disease, or something else? After stop word removal, stemming, and lemmatization, the bag-of-word representation resulted in a sparse matrix of size: 522,279 notes by 275,263 unique word tokens. We performed clustering with 30 clusters. The result was a meaningful clustering, with for example, top members of distinct clusters including: kidney ultrasounds, infant radiographs, and abdominal ultrasounds. The cluster list is shown in Figure 5 (right).

Conclusion
In this paper, we defined a deep clustering method to address this task of clustering high-dimensional sparse count data, as often arises in analysis of healthcare data and other fields. To do so, we described an autoencoder architecture that takes into account not just image recovery and clustering quality, but also geometric relationships in the data. We demonstrated that this approach works well through quantitative comparisons with other methods, as well as through detailed case studies on real healthcare data, including the development of an informatics tool that enables exploratory analysis of providers based on their clinical behaviors.

References
An Analytical Approach for Improving Patient-centric Delivery of Dialysis Services

Rosie Fleming, MM
t1, Daniel Gartner, PhD1,2, Rema Padman, PhD3 and Dafydd James, SPhT4
1School of Mathematics, Cardiff University, Cardiff, Wales, United Kingdom; 2National Health Service (NHS), Aneurin Bevan Continuous Improvement Unit (ABCi), Caerleon, United Kingdom; 3John Heinz III College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA; 4NHS, Abertawe Bro Morgannwg University Health Board, Renal Services, Swansea, United Kingdom

Abstract
In this paper, we report on the development of an analytical model and a decision support tool for meeting the complex challenge of scheduling dialysis patients. The tool has two optimization objectives: First, waiting times for the start of the dialysis after the patients' arrivals must be minimized. Second, the minimization of lateness after the scheduled finish time, which is relevant for transport services, are pursued. We model the problem as a mathematical program considering clinical pathways, a limited number of nurses managing the patients, and dialysis stations. Furthermore, information about patients' drop-off and pick-up time windows at/from the dialysis unit are considered. We develop a platform in Microsoft Excel and implement the analytical model using an Open Source optimization solver. A case study from a dialysis unit in the UK shows that a user can compute a schedule efficiently and the results provide useful information for patients, caregivers, clinicians and transport services.

Introduction
Recent research has reported an increasing prevalence of chronic kidney disease worldwide1. Patients who are in the end stage of chronic kidney disease (ESRD) rely on dialysis treatments to survive. Hemodialysis is a common approach to manage the condition. It is typically performed in a clinical setting three times a week for several hours, where the patient is connected to a machine via a vascular access. Patients seeking treatment in dialysis units have individual characteristics which can be distinguished on a variety of metrics such as blood flow rate, dialysate flow rate and composition, volume of fluid to be removed and size of dialyzer. Regular laboratory tests of the patient’s blood help the physicians to determine a suitable treatment plan. Patients may have many preferences, including the desire for short treatment times and preferred starting times during the day. Dialysis facilities, on the contrary, pursue planning the treatment efficiently by optimizing resource utilization for the best patient outcomes. Patient scheduling which is done in dialysis units world-wide, can be defined as assigning patients to scarce resources and time slots to maximize some objective2.

In this paper, we develop an analytical approach to deliver a patient-centric scheduling platform for dialysis patients. Individual patients' characteristics are taken into account, such as the patient’s clinical pathway, alongside the consideration of the availability of scarce resources that are required. We formulate the scheduling problem as a mathematical program which includes the scheduling decisions around the patients’ availability time windows. Using Microsoft Excel, a manager of a dialysis unit can specify parameters in our tool. The manager selects the waiting time objective and obtains feasible solutions based on the objective function and constraints which are implemented using the Open Source solver back-end of our tool. In doing so, patient schedules can be obtained effectively and efficiently such that waiting-times for patients can be minimized. Finally, the optimal schedules can be shared with patients, caregivers, clinicians and transport services.

The remainder of this paper is structured as follows. In the next section, we provide an overview of related patient scheduling work followed by the presentation of the analytical model, demonstration of the platform and discussion, and conclusions.
Related Work

Patient scheduling is the process of assigning individual patients and/or patients’ activities to time and/or healthcare resources on the operational decision level. In contrast, appointment scheduling defines a blueprint of patients’ appointments on a tactical level. While some reviews focus exclusively on patient scheduling, appointment scheduling problems have been reviewed by several authors. In this section, we position our paper in the relevant patient scheduling literature and focus on patient-related objectives as shown in Table 1.

<table>
<thead>
<tr>
<th>Minimize</th>
<th>8–14</th>
</tr>
</thead>
<tbody>
<tr>
<td>penalties</td>
<td>9–11, 15–30</td>
</tr>
<tr>
<td>waiting time of (prioritized) patients</td>
<td>31</td>
</tr>
<tr>
<td>welfare loss</td>
<td>32</td>
</tr>
<tr>
<td>number of night treatments</td>
<td>33</td>
</tr>
<tr>
<td>quality of life proxies</td>
<td></td>
</tr>
<tr>
<td>Maximize</td>
<td>19, 34–38</td>
</tr>
<tr>
<td># patients to be scheduled</td>
<td>17, 18, 22, 32, 39</td>
</tr>
<tr>
<td>patients’ satisfaction / preferences</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Patient-related objectives

The table reveals that most of the research focus has been on one single patient-related objective which is the minimization of patient waiting times which is similar to the objective of the collaborating dialysis unit. The difference of our work is, however, that we consider patient waiting times which occur before the start of the treatment and after the end of the treatment.

The analytical approach proposed in this paper can be categorized into and differentiated from the literature on patient scheduling as follows. One relevant paper focuses on hemodialysis scheduling but on a tactical decision level. The authors schedule patients’ treatments across several days and not within a day as we do. Furthermore, we provide a decision support tool that allows managers to accommodate patient availability and schedule patients more efficiently. In another relevant paper, therapy jobs are scheduled hospital-wide. The difference to our work is, again, that we provide a decision support tool that is based on an Open Source solver as compared to a commercial solver. Using Microsoft Excel and Open Source software increases the usability in the National Health Service because most of the computers have Microsoft Excel pre-installed, users are familiar with it and the Open Source package can be downloaded free and installed as a plugin. Another difference is that dialysis stations are considered as a scarce resource and patients’ clinical pathways consist of a setup, dialysis and a finish activity as our next section will reveal.

An Analytical Model

In what follows, we will introduce the parameters for our analytical model, the decision variables, objective function and constraints. Finally, an example schedule is given.

Parameters

Planning horizon, patients and activities

Let \( T := \{1, 2, \ldots, T\} \) be the set of 15-minute slots with planning horizon \( T \). In practice, we start the day at 7:00am and finish it at 11:00pm so that the planning horizon comes up to \( T = 64 \) periods enumerated using the set \( T := \{1, 2, \ldots, 64\} \). Dialysis patients are denoted by set \( \mathcal{P} \). \( A \) denotes the set of all clinical activities to be scheduled and \( A_p \subseteq A \) denotes the subset of activities for patient \( p \in \mathcal{P} \).

Hospital resources, capacity and demand

Nurses and stations have a capacity \( R_{nurse}^t \) and \( R_{station}^t \), respectively, in period \( t \in T \). For example, \( R_{nurse}^t = 1 \) means that 1 nurse is available between 7:00am and 7:15am. The demand of activity \( i \in A \) on nurses and stations is denoted by \( r_{nurse}^i \) and \( r_{station}^i \), respectively. This parameter will be used in the clinical pathways which are introduced next.
Clinical pathways

In our model, clinical pathways represent standardized, typically evidence-based health care processes as defined by van De Klundert et al. \(^{40}\). For more definitions and a literature review on clinical pathway modelling, see Aspland et al. \(^{41}\). We depict the clinical pathway of a patient as an activity-on-node graph in which the set of nodes represents the clinical activities. Weighted arcs represent minimum time lags between clinical activities and we write the activities’ capacity requirements below the nodes, see the legend in Figure 1.

Consider, for example, patients’ set of dialysis activities which are the setup, dialysis and the finish activity, depicted by \(\sigma_p, \delta_p,\) and \(\phi_p\), respectively shown in Figure 1. \(r_{\text{nurse}}^p = 1\) means that the patients’ setup activity requires 1 nurse who is busy with the patient setting her up on the dialysis machine.

Let \(E\) denote the set of all minimum time lags between clinical activities. A minimum time lag \((i, j) \in E\) of weight \(d^\text{min}_{i,j} \in \mathbb{Z}_{\geq 0}\) stipulates that activity \(j\) has to be scheduled at least \(d^\text{min}_{i,j}\) periods later than activity \(i\). Given the graph \((A, E)\), the earliest and latest start of activities, denoted by \(E_i\) and \(L_i\), respectively can be calculated using longest path methods (see, for example, \(^{42}\)). Let \(W_i := \{E_i, E_i + 1, \ldots, L_i\}\) denote the time window of activity \(i\). Once we have calculated the latest time slot \(L_{\phi_p}\) in which the dialysis activities can be scheduled, the index of the last period can be calculated by \(T = \max L_i\). Table 2 provides an overview of all parameters and decision variables. The latter will be introduced next.

*Figure 1: Clinical pathways for two dialysis patients*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>Set of activities</td>
</tr>
<tr>
<td>(A_p)</td>
<td>Set of activities corresponding to patient (p \in P)</td>
</tr>
<tr>
<td>(\delta_p)</td>
<td>Dialysis activity of patient (p) which excludes the setup and finish. They are modelled as separate activities.</td>
</tr>
<tr>
<td>(d^\text{min}_{i,j})</td>
<td>Minimum time lag for precedence relation ((i, j) \in E)</td>
</tr>
<tr>
<td>(E)</td>
<td>Set of precedence relations</td>
</tr>
<tr>
<td>(E_i)</td>
<td>Earliest period to schedule activity (i \in A)</td>
</tr>
<tr>
<td>(L_i)</td>
<td>Latest period to schedule activity (i \in A)</td>
</tr>
<tr>
<td>(P)</td>
<td>Set of patients</td>
</tr>
<tr>
<td>(\phi_p)</td>
<td>Dialysis finish activity of patient (p \in P)</td>
</tr>
<tr>
<td>(\phi)</td>
<td>Duration of activity (i)</td>
</tr>
<tr>
<td>(r_{\text{nurse}}^i)</td>
<td>Nurse demand by activity (i \in A)</td>
</tr>
<tr>
<td>(r_{\text{station}}^i)</td>
<td>Station demand by activity (i \in A)</td>
</tr>
<tr>
<td>(R_{\text{nurse}}^t)</td>
<td>Nurse capacity in period (t \in T) (e.g. 1 nurse available in period 1)</td>
</tr>
<tr>
<td>(R_{\text{station}}^t)</td>
<td>Station capacity in period (t \in T) (e.g. 1 station available in period 1)</td>
</tr>
</tbody>
</table>
Decision variables

We use binary activity-to-period assignment variables, a concept which has been used successfully in other scheduling literature, see\textsuperscript{43}. Accordingly,

\[
x_{i,t} = \begin{cases} 
1, & \text{if clinical activity } i \in A \text{ starts in period } t \in W_i \\
0, & \text{otherwise}.
\end{cases}
\]

Objective function

Having introduced all necessary parameters and decision variables, the objective functions are given by Equations (1) and (2).

\[
\begin{align*}
\text{Minimize } & z = \max_{p \in P} \sum_{t \in W_p} t \cdot x_{\sigma_p,t} \\
\text{Minimize } & z = \max_{p \in P} \sum_{t \in W_p} t \cdot x_{\phi_p,t}
\end{align*}
\]

Objective function (1) minimizes the maximum waiting time for patients to start the dialysis session as follows: Each patient \( p \) has a time window to start the session which is defined for the individual’s dialysis start activity \( \sigma_p \) (see the patients’ clinical pathways shown in Figure 1). Now, one decision variable \( x_{\sigma_p,t} \) is equal to 1, we multiply the time point \( t \) which leads to the scheduled start times across all patients. Now, the maximum value of this start time vector is minimized. Similarly, for objective function (2), we minimize the maximum scheduled finish time of the treatments.

Constraints

In what follows, we add constraints to our model which we break down by clinical pathways, nurse and machine constraints.

Clinical pathways, nurse and machine constraints

Constraints (3) use the information from the clinical pathways defined earlier and ensure that minimum time lags between all consecutive activities are guaranteed.

\[
\sum_{t \in W_j} t \cdot x_{j,t} \geq \sum_{t \in W_i} t \cdot x_{i,t} + \delta_{i,j}^{\min} \quad \forall p \in \mathcal{P}, (i, j) \in \mathcal{E}_p
\]

Nurse constraints (4) ensure that the demand for nurses does not exceed the nurse capacity.
\[
\sum_{i \in A} \sum_{t \in W_i} r_{i,nurse}^{\min(L_i,t)} \sum_{\tau = \max(E_i,t-p_i)+1}^{\min(L_i,t)} x_{i,\tau} \leq R_{t,nurse}^{nurse} \quad \forall t \in T
\] (4)

Constraints (5) ensure that the demand for the stations does not exceed the station capacity.

\[
\sum_{i \in A} \sum_{t \in W_i} r_{i,station}^{\min(L_i,t)} \sum_{\tau = \max(E_i,t-p_i)+1}^{\min(L_i,t)} x_{i,\tau} \leq R_{t,station}^{station} \quad \forall t \in T
\] (5)

Constraints (6) ensure that each activity is scheduled exactly once.

\[
\sum_{t \in W_i} x_{i,t} = 1 \quad \forall i \in A
\] (6)

Variable definitions and their domains are given by (7).

\[
x_{i,t} \in \{0, 1\} \quad \forall i \in A, t \in W_i
\] (7)

**Example**

Table 3 shows a station and nurse allocation example based on the clinical pathways defined in Figure 1. “–” means that the variables are not defined in these periods because they are outside the activities’ time windows. We assume that the processing times of the patients’ start activities are 1, the finish activities require two time periods and the durations of the dialysis (\(\delta_p\)) take 4 and 1 period for patients 1 and 2 respectively. Naturally, the durations are longer in reality as the next section will reveal.

<table>
<thead>
<tr>
<th>(t \in T)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_{\sigma_1,t})</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>(x_{\delta_1,t})</td>
<td>–</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>(x_{\phi_1,t})</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(x_{\sigma_2,t})</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>(x_{\delta_2,t})</td>
<td>–</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>(x_{\phi_2,t})</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>(\sum_{i \in A} \sum_{t \in W_i} r_{i,station}^{\min(L_i,t)} \sum_{\tau = \max(E_i,t-p_i)+1}^{\min(L_i,t)} x_{i,\tau})</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(\sum_{i \in A} \sum_{t \in W_i} r_{i,nurse}^{\min(L_i,t)} \sum_{\tau = \max(E_i,t-p_i)+1}^{\min(L_i,t)} x_{i,\tau})</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

The example reveals that the first activity is scheduled in period \(t = 1\) which means that the nurse is allocated in the same period. In the next period \((t = 2)\) the second patient’s first activity \((\sigma_2, \text{see Figure 1})\) is scheduled which allocates the nurse in the second period \((t = 2)\) as can be seen in the second to last row. In period \(t = 3\), the nurse is not assigned to tasks that involve the connection and disconnection of patients to/from machines.
The last row shows the demand profile for the capacity requirement from the dialysis stations. As can be seen, one station is allocated in period $t = 1$, followed by an allocation of two stations in periods $t = 2, 3, \ldots, 6$. As only one patient is on the station at period $t = 7$, the demand profile goes down to 1.

**Decision Support Tool**

We created a decision support tool in Microsoft Excel. The tool is broken down into a parameters tab which is shown in Figure 2, the solver tab shown in Figure 3, and a solution tab shown in Figure 4.

Figure 2 shows the user interface where the manager can input each patient, along with their treatment duration and arrival time on the left-hand side. Sometimes, a station may be unavailable because of maintenance operations going on for the dialysis machine. Accordingly, the user can parametrize the station availability by using the “station availability table” on the right-hand side. The nurses’ availability and the cleaning time for the machines can also be inputted.

<table>
<thead>
<tr>
<th>Patient Name</th>
<th>Patient ID</th>
<th>Treatment Duration (hours)</th>
<th>Ready Time (in hours after 7am)</th>
<th>Station Availability</th>
<th>Start:</th>
<th>End:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omar Scruggs</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Fiona Swingle</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Sharee Bax</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Debrah Wallford</td>
<td>4</td>
<td>3.5</td>
<td>0</td>
<td>4</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Valentine Manor</td>
<td>5</td>
<td>3.5</td>
<td>0.25</td>
<td>5</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Carly Ploss</td>
<td>6</td>
<td>3.5</td>
<td>0.25</td>
<td>6</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Michale Wertz</td>
<td>7</td>
<td>3.5</td>
<td>0.25</td>
<td>7</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Shanilay Wasser</td>
<td>8</td>
<td>3.5</td>
<td>0.25</td>
<td>8</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Aviane Grauer</td>
<td>9</td>
<td>3.5</td>
<td>0.5</td>
<td>9</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Kristian Hankerson</td>
<td>10</td>
<td>3.5</td>
<td>0.5</td>
<td>10</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Kerstin Niser</td>
<td>11</td>
<td>4</td>
<td>0.5</td>
<td>11</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Rayna Pentland</td>
<td>12</td>
<td>4</td>
<td>0.5</td>
<td>12</td>
<td>07:00</td>
<td>23:00</td>
</tr>
<tr>
<td>Winston Tooper</td>
<td>13</td>
<td>4</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emilio Porcelli</td>
<td>14</td>
<td>4</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dorine Aman</td>
<td>15</td>
<td>4</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arden Aiken</td>
<td>16</td>
<td>4</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dong Eakinns</td>
<td>17</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gloria Pateck</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tami Pompa</td>
<td>19</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zofia Nowsky</td>
<td>20</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audrea Whitel</td>
<td>21</td>
<td>4</td>
<td>1.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosaria Bacher</td>
<td>22</td>
<td>4</td>
<td>1.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chere Eye</td>
<td>23</td>
<td>4</td>
<td>1.25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2:** Parameters for scheduling the patients in the Dialysis Unit. All patients’ names are synthetic.

In Figure 3, the user can choose which objective function they wish to optimize when solving the scheduling problem. As mentioned in the modelling section, we have two patient lateness objectives, the first being the time they wait after arrival before starting their treatment, and the latter being the time a relative, Welsh ambulance or a private taxi service has to wait for the treatment to finish. Once the user selects the objective function, OpenSolver will use the inputted values and constraints to create a schedule which is then displayed to the clinician using metrics such as activity start times and lateness.
Figure 3: Solving the problem of scheduling the patients in the Dialysis Unit

Figure 4 shows the output of the tool in a Gantt chart. For each station, we can see which patients have been scheduled at time \( t \). The black bars relate to the patients’ start activity \( \sigma_p \) and finish activity \( \phi_p \) (introduced in Figure 1). Recall that these activities require nurses in addition to the stations. The coloured bars relate to activity \( \delta_p \), where only the station is required for the actual dialysis activity \( \delta_p \).

<table>
<thead>
<tr>
<th>Station 1</th>
<th>Station 2</th>
<th>Station 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow Time</td>
<td>4.5</td>
<td>9</td>
</tr>
<tr>
<td>Lateness calculation 1</td>
<td>0.5</td>
<td>4.25</td>
</tr>
<tr>
<td>Tardiness</td>
<td>0.5</td>
<td>4.25</td>
</tr>
<tr>
<td>Flow Time</td>
<td>4.8</td>
<td>12.75</td>
</tr>
</tbody>
</table>

Discussion

In several workshops we discussed the usability and acceptability of the platform with managers from the National Health Service. One major feedback to improve the usability of the platform was that we have to link our tool with the clinical information system which operates on a separate database. However, using an ODBC connection, we will be able to link Microsoft Excel with the patient data in the collaborating dialysis unit.

The usage of our platform will have several implications and benefits for different stakeholders: Using our tool, patients can be communicated an expected dialysis start time which may reduce perceived waiting time. Nurses will be able to plan their breaks during the day which is very difficult at the moment because of the workload imposed by the unfavorable schedules: Currently, nurses struggle to take breaks during midday because patients who started the treatments in the morning finish at midday and need to be taken care of. In parallel, new patients who arrive at midday have to be setup on the machines and delays may impose overtime at the end of the nurses’ shifts. Another benefit of our structured approach is that transport services and patients’ caregivers can be informed about the scheduled dialysis.
finish times. Using information about the patients’ home addresses, transport services may aggregate trips for patients and provide more efficient services.

The mathematical model has been formulated as a deterministic problem which means that uncertainty is not taken into account at this stage. Uncertainty may happen in patients’ no shows, late arrivals and uncertain dialysis durations. Using information from the past, however, machine learning algorithms may be used to accurately predict no-shows which then can be incorporated into our tool. Alternatively, a rolling-horizon procedure may be used to take into account variation during the execution of the schedule.46

Conclusions

In this paper we have presented an analytical model and a decision support tool for the problem of scheduling dialysis patients in a dialysis unit in the UK. One objective is to minimize the maximum waiting time for patients to start the dialysis session. The second objective minimizes the maximum scheduled finish time of all treatments. In doing so, the model avoids nurses’ overtime at the end of the day. Using data from a hospital we demonstrated the effectiveness of our approach and showed the solution output using a Gantt chart. This helps the dialysis unit to find out the optimal sequence of patients on the different dialysis stations.

Future work will include further patient related objective functions including patient preferences, clinician guidelines and targets, and also taking into account resource-related measures such as utilization maximization. Furthermore, we will evaluate the importance of each of the different objectives and incorporate the result into a multi-criteria optimization approach. Also, our aim is to quantify the effectiveness of the approach in practice.

Acknowledgments

This research was supported by the Welsh Health Hack 2017 in collaboration with the Welsh Government, Bevan Commission, ABCi and the Wales Deanery. Additional funding was received from the Bevan Commission’s Health Technology Exemplar scheme. Also, the Data Innovation Research Institute at Cardiff University provided Seedcorn Funding to support the project and to facilitate new collaborations.

References


Engaging Pharmacists to Crowdsourced a Fine-grained Medication Risk Scale: An Initial Measurement Study Using Paired Comparisons of Medications

Allen J. Flynn, PharmD, PhD, Greg Farris, BS, George Meng, MSEE, Jack Allan, Sara Kurosu, Natalie Lampa, Koki Sasagawa, BS
Medical School, University of Michigan, Ann Arbor, MI

Abstract

A coarse classification of medications into two risk categories, one for high-risk medications and one for all others, allows people to focus safety improvement work on medications that carry the highest risks of harm. However, such coarse categorization does not distinguish the relative risk of harm for the majority of medications. To begin to develop a more fine-grained measurement scale for the relative risk of harm spanning many medications, we performed an experiment with 18 practicing pharmacists. Each pharmacist-participant made 210 paired comparisons of 21 commonly prescribed medications to reveal a subjective scale of perceived medication worrisomeness (PMW). Statistical analyses of their collective judgments of medication pairs differentiated five levels of PMW. This study illuminates one path towards a fine-grained medication risk scale based on PMW. It also shows how the method of paired comparisons can be used to remotely crowdsourced expert knowledge in support of learning health systems.

Introduction

Current approaches to medication safety typically focus on mitigating risks from using high-alert medications. According to the Institute for Safe Medication Practices (ISMP), a high-alert medication is “a drug that bears a heightened risk of causing significant patient harm when it is used in error”\(^1,2\). Based on evidence about medication errors and adverse drug events, ISMP enumerates individual high-alert medications. One example is promethazine injection, which has a low pH and therefore causes severe injuries in cases of extravasation\(^3\). ISMP also specifies that certain therapeutic drug classes are comprised only of high-alert medications, e.g., the opioid therapeutic drug class\(^4\).

It is useful to develop policies and procedures for handling high-alert medications like promethazine and opioids\(^4\). Pharmacists often lead efforts to develop such policies and procedures within provider organizations\(^5\). However, the current binary approach to medication risk categorization, where medications are classified either as high-alert or NOT high-alert, is too coarse to meet growing risk assessment needs in practice. With only these two risk categories, it is impossible to assess the relative risk of harm (RRH) of a specific medication in comparison to others. Also, the current two-category approach to risk categorization is not suitable for determining the composite relative risk of harm (cRRH) of a list of prescribed medications comprising a patient medication regimen. However, multiple studies suggest that relative risk scores for individual prescriptions and whole medication regimens are needed in practice to direct attention to potential problems and to allocate provider effort in more optimal ways\(^6–8\).

The broadest goal of this work is to improve medication safety systematically by first developing and then applying a reliable, valid, and fine-grained measurement scale to accurately indicate to prescribers and patients the RRH for every prescribed medication. In the absence of such a scale, pharmacists are devising makeshift scoring tools within electronic health record systems (EHRs) to try and identify patients at the greatest risk for adverse drug events\(^9\). This paper reports an initial experiment to realize a new approach to creating a fine-grained medication RRH measurement scale. We show that a crowdsourced scale can be built using paired comparisons and remote knowledge acquisition.

Background and Significance

Besides scoring the relative risk of individual prescriptions and whole medication regimens, there are a growing number of other uses for a reliable, numeric RRH measurement scale for medications\(^9\). For example, to safely and incrementally implement artificial intelligence (AI) in pharmacy practice, to uphold safety there is a need to initiate AI solutions in a limited way beginning with those medications with a low or very low RRH\(^10,11\). Also, with a more fine-grained medication RRH scale, it would be possible to use risk scores arising from applying the scale as inputs to try and improve many holistic patient risk models, such as hospital readmission risk models\(^12\).

Ideally, large quantities of accurate and current data about the actual harm caused by medications would be used to derive a reliable and valid measure of RRH for all medications. Unfortunately, high-quality medication error and adverse drug event data are not available in sufficient quantities to build a comprehensive RRH measurement scale in this way. Estimates indicate that less than 20 percent of actual medication errors and near misses are reported\(^13\). Such
sparse and incomplete data only support the identification of the most harmful, high-alert medications. This paper explores the feasibility of constructing a fine-grained medication RRH measurement scale with data about perceived medication worrisomeness (PMW) collected from practicing pharmacists using the method of paired comparisons.

There are several reasons to believe that pharmacists may develop an expert sense of the relative risk of harm for a wide range of medications as they practice pharmacy. First, pharmacists are accountable for therapeutic and other outcomes stemming from the use of a vast array of drugs. Second, pharmacists’ interest in understanding the full breadth of the medication armamentarium distinguishes their practice from specialty medical and nursing practices. Third, pharmacists often develop risk-based medication safety policies and guidelines and apply them in practice. Fourth, pharmacists are exposed to adverse drug events, likely adding to their insight about medication risks.

If pharmacists do develop a sense of RRH of medications through practice, then the method of paired comparisons could offer an efficient and effective way to acquire this expert knowledge from them. The analytic method of using paired comparisons has at least a 90-year history. Paired comparisons are used in many scientific fields to establish reliable and valid rating scales. The method of paired comparisons is especially useful when no natural measurement scale is available, which is the case for a scale to measure the RRH of many medications. For this study, we chose to infer RRH from perceived medication worrisomeness (PMW) while recognizing that the RRH of medications is most likely a complex multidimensional concept deserving of further, more detailed examination.

The method of paired comparisons provides a way to rank multiple items. Via experiments, paired comparisons enable us to estimate the probabilities of true ratings of multiple items along a continuum of interest. First, the set of all pairs of items are defined, then participants are asked to judge items against one another head-to-head in a manner that reveals a subjective continuum. Once these judgments are made, the paired comparison data provided by participants can be fit to specialized linear models, like the Bradley-Terry model.

In the Bradley-Terry model, paired comparison data are represented in terms of initial probability estimates where

\[
\text{Probability (item } i > \text{ item } j) = \frac{\alpha_i}{\alpha_i + \alpha_j},
\]

\(i \neq j\) and \(\alpha_i\) and \(\alpha_j\) are positive-valued parameters associated with items \(i\) and \(j\), for each of the paired comparisons. These data are then combined for an entire experiment and fitted to a hyperbolic secant probability distribution function. By fitting these data this way, researchers can compute maximum likelihood estimates of the true probability of each item’s performance, along with a corresponding degree of uncertainty. In light of their uncertainties, the maximum likelihood estimates of the true probabilities can then be rank ordered on a single continuum of interest.

We designed and executed a paired comparison experiment. It involved repeated paired comparisons of perceived medication worrisomeness made by pharmacists. We did this to demonstrate, in a preliminary way, the feasibility of using this type of experiment to reveal a comprehensive, fine-grained RRH scale for medications by remotely crowdsourcing pharmacists’ expert knowledge of medication concerns using a novel online data collection tool.

Medications generally have different risk profiles. It follows that we must manage the risk of harm from medications on a medication-specific basis. We recognize that in many but not all cases, patient-specific factors mediate or moderate the RRH of medications. However, just as we think of the likelihood of medication side effects for all people, we believe it is reasonable and equally necessary to focus on the general risks of harm from medications for everyone.

To better allocate limited clinician resources and mitigate the risk of harm from using medications, a fine-grained measure of the RRH of prescribed medications is needed. To our knowledge, a broadly conceived, reliable, valid, fine-grained measure of RRH for medications does not exist. This study is significant because it begins to address this unmet need in a rigorous, systematic way. We report an initial step towards a comprehensive measure for assessing whether various prescribed medications have RRHs that are very low, low, moderate, high, or very high. To improve safety and spare provider work-time, we hope that in the future such a measure will direct people’s attention toward a wider array of significant medication risks and away from other insignificant medication issues.

Research Questions

For this study, we investigated three research questions. The first two questions, RQ1 and RQ2, represent tests of the major hypothesis proposed by Bradley and Terry for all paired comparison experiments. To answer RQ1 and RQ2, we test whether or not statistically significant differences from random selection exist to rank medications on a subjective scale of perceived medication worrisomeness. The third question, RQ3, addresses how to scale-up an experiment like this one to achieve a comprehensive fine-grained RRH scale for hundreds of medications.
**RQ1.** To what degree do a group of practicing pharmacists collectively perceive differences in worrisomeness among single medications in a collection of 21 commonly prescribed medications?

**RQ2.** To what degree do a group of practicing pharmacists collectively indicate differences in worrisomeness among individual therapeutic categories of medications in a collection of 7 common therapeutic categories?

**RQ3.** What are some key requirements for using the method of paired comparisons to crowdsource a comprehensive relative risk of harm scale spanning hundreds of medications?

**Methods**

We begin with a relevant but fictional worked example of our analytic method which comes primarily from the work of Bradley and Terry\textsuperscript{17,18}. After the worked example, we describe more details of our specific experimental methods.

**Worked Example of the Analytic Method of Paired Comparisons Applied to Fictitious Medication Comparisons**

Consider the following fictional experiment where perceived medication worrisomeness (PMW) for 4 medication items is examined using the method of paired comparisons. The 4 medications are aspirin, acetaminophen, ibuprofen, and ketorolac. They are non-opioid analgesics with varying risk profiles\textsuperscript{21}. The set of all pairs of these 4 medications has a cardinality of \(\binom{4}{2}\), or six. These six pairings appear in rows 1 to 6, columns II and III, of Table 1 below.

Now, imagine that 20 pharmacists compare the six pairs of medications in Table 1 head-to-head by indicating, independently for each pair, which medication in the pair would cause them to worry most for a patient who takes it. From these 20 repetitions, assume the fictional but realistic data shown in columns IV and V of Table 1 arise. The data in the first-row show that 8 fictitious pharmacists perceived aspirin (1\textsuperscript{st} in pair) to be more worrisome than acetaminophen while 12 other pharmacists perceived acetaminophen (2\textsuperscript{nd} in pair) to be more worrisome than aspirin.

**Table 1.** Fictional data for a worked example of the method of paired comparisons applied to medication choices

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1\textsuperscript{st} in pair</td>
<td>2\textsuperscript{nd} in pair</td>
<td>1\textsuperscript{st} more worrisome</td>
<td>2\textsuperscript{nd} more worrisome</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>aspirin</td>
<td>acetaminophen</td>
<td>8</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>aspirin</td>
<td>ibuprofen</td>
<td>15</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>aspirin</td>
<td>ketorolac</td>
<td>2</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>acetaminophen</td>
<td>ibuprofen</td>
<td>17</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>acetaminophen</td>
<td>ketorolac</td>
<td>3</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ibuprofen</td>
<td>ketorolac</td>
<td>0</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

With the data in Columns IV and V of Table 1, Equation 1 above can be used to generate estimates of the true probabilities describing how pharmacists perceive the relative worrisomeness of medications in each pair. Then, by fitting the data in Columns IV and V to an appropriate statistical model, it is possible to test whether or not any of these six medications differ from one another in overall perceived medication worrisomeness (PMW).

To rank order the four medications, \(M_1 \ldots M_4\), for PMW after arriving at parameter estimates with standard errors for each one, the data in Table 1, columns IV and V can be fitted to a statistical model for combining paired comparison data. One commonly used model for this comes from Bradley and Terry\textsuperscript{17,18}. Here is an expression of their model in the context of this worked example:

\[
\text{logit} \left[ \text{Probability} \left( M_i \text{ is more worrisome than } M_j \right) \right] = \lambda_i - \lambda_j ,
\]

where \(\lambda_i = \log(\alpha_i)\) for all \(i\), and \(\lambda_j = \log(\alpha_j)\) for all \(j\), for the entire experiment. Assuming all comparisons are independent, parameters \(\{\lambda_i, \lambda_j\}\) for the overall PMW of each medication can be estimated by maximum likelihood.

To estimate the parameters \(\lambda_1, \ldots, \lambda_4\), which correspond to \(M_1, \ldots, M_4\), a system of two maximum likelihood equations can be solved iteratively\textsuperscript{17}. To solve these equations, we used the BradleyTerry2 package for the R statistical computing platform\textsuperscript{19}. What results from this model-fitting procedure are maximum likelihood estimates (MLEs) with standard errors (SEs) for the PMW of aspirin, ibuprofen, and ketorolac. The MLE and SE for acetaminophen set to zero by convention (Table 2, Columns II and III). In Column V of Table 2, we then use Firth and de Menezes method of calculating quasi-standard errors for these MLEs to allow statistical inferences to be made about the degree of difference in PMW between contrasting pairs of medications\textsuperscript{22}.
Table 2. Fictional results for a worked example of the method of paired comparisons for 4 medications. Column III shows maximum likelihood estimates for perceived medication worrisomeness from the Bradley-Terry model.

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III MLE</th>
<th>IV SE</th>
<th>V Quasi SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ketorolac</td>
<td>1.86 (highest)</td>
<td>0.50</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>acetaminophen</td>
<td>0.00</td>
<td>0.00</td>
<td>0.27</td>
</tr>
<tr>
<td>3</td>
<td>aspirin</td>
<td>-0.47</td>
<td>0.37</td>
<td>0.26</td>
</tr>
<tr>
<td>4</td>
<td>ibuprofen</td>
<td>-1.73 (lowest)</td>
<td>0.45</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Generating results like the fictional results in Table 2 above is a goal of this study. At this point, we have completed our review of a fictitious worked example detailing the analytic and statistical methods for this study.

Medication and Therapeutic Category Selection

For this study, we used the two criteria listed directly below to select a total of 21 medications for comparison. The five medications marked with a diamond (♦) are designated as high-alert medications by ISMP.

1. Each medication appears on the Top 200 most commonly prescribed drugs list for 201823
2. Each medication is a member of one of these seven therapeutic drug categories or drug classes
   a. Antibiotics (amoxicillin, azithromycin, ciprofloxacin)
   b. Anticoagulants (apixaban♦, rivaroxaban♦, warfarin♦)
   c. Antidepressants (amitriptyline, citalopram, sertraline)
   d. Antidiabetics (glyburide♦, liraglutide, metformin♦)
   e. Antihypertensives (amlodipine, lisinopril, losartan)
   f. Benzodiazepines (alprazolam, clonazepam, lorazepam)
   g. Non-steroidal anti-inflammatories (aspirin, ibuprofen, naproxen)

To reach our complement of 21 medications, those that are listed above in parentheses, we selected three medications from seven different therapeutic drug categories. Only very commonly prescribed medications were included in an effort to ensure that pharmacists would be very familiar with all 21 medications. Settling on 21 as the total number of medications gave rise to a total of \( \binom{21}{2} \) or 210 pairs of medications for comparison by each pharmacist-participant. To show that this number of comparisons is workable, we ran several tests and confirmed that pharmacist-participants can comfortably make 210 paired medication comparisons in less than an hour.

We intentionally included medications for acute problems (e.g., antibiotics) and others for chronic diseases (e.g., antihypertensives). We included controlled substances (i.e., benzodiazepines) and non-controlled substances. Most of the medications come in oral tablets or capsules, but we added one injectable medication (liraglutide). Also, we included concerning medications associated with many adverse drug events (e.g., anticoagulants24) but also generally safe over-the-counter medications (non-steroidal anti-inflammatory drugs called NSAIDs).

Rationale for Not Permitting Ties When Pharmacist-Participants Make Paired Comparisons

We did not permit pharmacist-participants to declare ties in PMW when making paired comparisons. Our rationale for not permitting ties is based on the assumption that no two medications have precisely the same risk profile or carry an identical risk of harm. To check this assumption, we collected timestamps for each paired comparison choice made serially by pharmacist-participants and used this timestamp information to look for any paired comparisons that were difficult or very difficult for the majority of pharmacist-participants to make.

Determination of the Number of Pharmacist-Participants to Recruit

Using the method for estimating experiment size from the second edition of David’s 1988 textbook on paired comparisons (p. 109)25, we sought to determine an appropriate number of pharmacist-participants to recruit. Recall that we have a fixed number of 21 medication items (or treatments, t) to be compared and ranked. David indicates that, for \( t=21 \), to ensure at least a predetermined probability of 0.95 for the selection of the medication perceived to be most worrisome, assuming a minimum parameter difference of between 0.1 and 0.15 between the medication perceived to be most worrisome and the next most worrisome medication(s), 10 to 20 subjects must make all 210 paired comparisons. On this basis, we sought to recruit between 10 and 20 pharmacist-participants for this study.
Recruitment of Pharmacist-Participants

For this study, we recruited pharmacist-participants who met the following two inclusion criteria.

1. The pharmacist-participant is licensed and registered to practice pharmacy in at least one U.S. state
2. As part of their current work role, the pharmacist-participant performs some direct patient care duties

To recruit pharmacist-participants who met the two criteria above, we executed two non-probability sampling strategies. Our primary recruitment strategy was referral or snowball sampling. We sent invitation messages by e-mail to 20 hospital, ambulatory care, and community pharmacists asking them to participate and to invite other pharmacists they know to participate. We deliberately sent these messages to individuals in several states, including Michigan, Ohio, Indiana, Illinois, and Connecticut. Our secondary recruitment strategy was to advertise on the Michigan Pharmacists Association’s public LinkedIn account. We posted three paid messages inviting pharmacists to participate. When pharmacists responded, we shared more information, including a consent form, and offered them a $50 gift card in exchange for an hour of their time spent as a study participant.

Design, Development, and Testing of CrowdSort – An Online Paired Comparison Data Collection Tool

For this study, we developed an online web application tool, or web app, to collect paired comparison data from remote pharmacists (Figure 1). We call this web app CrowdSort because it enables people to participate in sorting or ranking using the method of paired comparisons. With CrowdSort, we collected data when participants joined us remotely from their worksites around the U.S. for individually-scheduled, hour-long conference calls.

CrowdSort’s source code is available at github.com/kgrid-demos/crowdsort. Three team members led CrowdSort development (AF, GM, JA). For the front end, several team members (JA, SK, NL, KS) used HTML, CSS, Vue.js, and JavaScript to establish CrowdSort’s user interface. For the back end, three team members (GF, GM, SK) deployed CrowdSort on a cloud server at Heroku (heroku.com) and then used the mLab Heroku add-on to establish a persistent connection to a database operating on the Heroku platform. Once deployed in this way, the CrowdSort web app posted each paired comparison to our cloud database with a date and timestamp. Three team members (AF, JA, KS) developed and cross-checked two software tools for transforming the raw paired comparison data into a format suitable for analysis using the BradleyTerry2 package for the R statistical computing platform.

We designed CrowdSort’s user interface so that it would be simple to use. Our user interface design process began by examining the user interfaces of the comparison-making websites either.io and rrrather.com. With the designs of these two existing websites in mind, we created CrowdSort’s similar user interface (Figure 1).

For users to complete each paired comparison using CrowdSort, they simply click on one of the two text options appearing in the two onscreen boxes (Figure 1). These two boxes, highlighted by a shadow effect, are the chief feature of this user interface. For that reason, they have different colors and appear in the middle of the screen.

Figure 1. View of CrowdSort web application for making paired comparisons online with a web browser
As an example of a paired comparison, in Figure 1 above, a single comparison is portrayed between the antibiotic ciprofloxacin in the blue box on the left and the antidiabetic drug metformin in the brown box on the right.

After the user clicks on one box or the other to make their choice, a new pair of medications is displayed until all 210 paired comparisons have been made. This first version of CrowdSort, which is accessible online at crowdsort.herokuapp.com, is designed NOT to allow users to indicate ties in perceived medication worrisomeness.

To mitigate order effects, whenever a web browser reloads CrowdSort, both the order of the 210 pairs and their left-right position onscreen are randomly shuffled. To mitigate effects of using colors to distinguish the two onscreen boxes visually, we chose the colors brown and blue because they are NOT safety related colors, like red, orange, yellow, or green. The two colors also randomly shift between left and right as the user makes paired comparisons.

To test CrowdSort, four research team members worked to develop, document, and run tests of this system, uncovering and fixing software defects in the process (JA, SK, NL, KS). Then, to confirm CrowdSort was ready to support remote data collection, we conducted two trial runs with University of Michigan student pharmacists. Because these two trial runs were both successful, data collection activities commenced.

Prompt to elicit pharmacists’ thoughts about perceived medication worrisomeness

For this study, we gave careful consideration to the prompt and instructions for pharmacist-participants who performed the comparisons of 210 pairs of medications. We prompted them to select which medication in each pair would cause them to worry more than the other if the medication was indicated and prescribed for one of their patients (Figure 1). This prompt was devised to elicit pharmacists’ initial reactions to each pair of medications in a way that does not suggest a correct answer. As much as possible, we wanted to avoid having the task of making comparisons feel like taking a test or exam. We deliberately avoided the term “risk” in the prompt because, unlike “worry”, which is subjective, “risk” could elicit objective or legalistic meanings in the minds of participants. We assume that the concepts of perceived medication worrisomeness and perceived medication risk are closely and positively correlated.

Results

This section begins with several overall results followed by results which address each research question, RQ1-RQ3.

A total of 18 pharmacist-participants from seven different states were recruited for this study. Eight men and 10 women were included. Seven pharmacist-participants were community pharmacists and 11 were hospital pharmacists. All 18 pharmacist-participants completed the data collection process by using the CrowdSort web app to perform 210 paired comparisons. Their work resulted in a dataset with 3,780 paired comparisons of PMW, including 360 comparisons for each of the 21 medications. After removing all comparisons of drugs in the same therapeutic category, we found that pharmacist-participants indirectly made 3,402 paired comparisons of the seven therapeutic categories listed above.

We analyzed the time required for pharmacist-participants to make 210 paired medication comparisons with the CrowdSort web app. On average, they completed this task in 16 ± 8 minutes (Range: 9 to 41 minutes). To detect whether some comparisons might have been more difficult than others, we used the timestamps for each comparison to look for paired comparisons that required pharmacist-participants three or five times more than their average comparison-making interval. No pairing required significantly more time than average for the majority of participants.

RQ1. To what degree do a group of practicing pharmacists collectively perceive differences in worrisomeness among single medications in a collection of 21 commonly prescribed medications?

Results relevant to answering RQ1 are illustrated and summarized in Figure 2 below. We found that our small crowd of 18 practicing pharmacists collectively perceive significant differences in perceived medication worrisomeness (PMW) among the 21 medications compared. The anticoagulant warfarin is perceived as most worrisome overall. It was judged more worrisome in 346/360 (96%) of its head-to-head pairings. The hypertension medication amlodipine and the antibiotic amoxicillin were least worrisome overall. These two medications were judged more worrisome in only 48/360 (13%) and 58/360 (16%) of their head-to-head pairings, respectively. Using quasi standard errors to enable medication-by-medication comparisons, our results reveal five emerging PMW groups (Figure 2). Between the groups of medications perceived to be least worrisome and most worrisome, three intermediate groups can be seen.

As noted, warfarin♦, which is marked with a diamond because it is an ISMP high-alert medication, has the highest estimated PMW. Next, in the High Intermediate group, two other high-alert anticoagulants, apixaban♦ and rivaroxaban♦, appear (Figure 2). The Intermediate group includes all 3 benzodiazepines (alprazolam, clonazepam, lorazepam), an antidepressant (amitriptyline), an antibiotic (ciprofloxacin), and a high-alert antidiabetic (glyburide♦).
Figure 2. PMW estimates with Quasi Standard Errors for 21 medications abbreviated on the X-axis. Diamonds (♦) indicate ISMP high-alert medications. Select estimates are noted. Percentages are the empirical proportion of pairings when amiodipine, amoxicillin and warfarin were judged more worrisome.

The Low Intermediate group includes all 3 NSAIDS (aspirin, ibuprofen, naproxen), an antibiotic (azithromycin), two antidepressants (citalopram, sertraline), two antidiabetic drugs (liraglutide, metforminic), and two antihypertensives (lisinopril, losartan). Finally, amoxicillin and amiodipine have the lowest estimated PMW scores (Figure 2).

We assessed goodness of fit and found that our medication paired comparison data fit the Bradley-Terry model well. To do this, we computed standardized residuals for each of the 210 paired comparisons. Upon inspection, we found the distribution of these residuals to be platykurtic and slightly skewed compared to the normal distribution (Kurtosis = 2; Skewness = 0.25; W/S = 7.5, n=210). Besides, we found that 171 of 210 (81%) of the fitted probabilities fit within less than one standard deviation from the Standardized Residual Mean (SRM), while 35 (17%) fell between 1 and 2 standard deviations of the SRM and four (2%) fell beyond 2 standard deviations from the SRM.

RQ2. To what degree do a group of practicing pharmacists collectively indicate differences in worrisomeness among individual therapeutic categories of medications in a collection of 7 common therapeutic categories?

Results relevant to this question are illustrated and summarized in Figure 3 below. By making 210 paired comparisons of medications associated with the seven therapeutic categories, participant-pharmacists indirectly indicated that significant differences exist in their minds regarding the relative worrisomeness of these therapeutic drug categories.

As shown below in Figure 3, we infer from the paired comparison data for medications that the therapeutic category of medications perceived to be most worrisome is the anticoagulants category. Pharmacist-participants indicated the anticoagulant in a pair was more worrisome than medications from all other classes 93% of the time. The therapeutic category evidently perceived to be least worrisome was antihypertensives. Pharmacist-participants indicated that the antihypertensive medication in a pair was more worrisome than medications from other therapeutic classes only 21% of the time. Using quasi standard errors to enable category-by-category comparisons, our results reveal a total of four distinct, emerging groups of inferred therapeutic category worrisomeness (ITCW).

As a class, anticoagulants have the highest estimated PMW. Next, in the High Intermediate group, the benzodiazepine therapeutic category appears by itself (Figure 3). The Low Intermediate group includes antibiotics, antidepressants, antidiabetics, and NSAIDs. The least worrisome therapeutic category in these results is the antihypertensives category.

We assessed goodness of fit and found that our inferred paired comparison data for the seven therapeutic categories fit the Bradley-Terry model moderately well. To check goodness of fit we computed standardized residuals for the (2) or 21 comparisons. Upon inspection, we found the distribution of this low number of 21 residuals to be very platykurtic and a little skewed compared to the normal distribution (Kurtosis = 0.2; Skewness = -0.5; W/S = 3.9, n=21). However, we also found that all 21 of the fitted probability values from using the Bradley-Terry model for the paired comparisons for the therapeutic categories fit within less than one standard deviation from the Standardized Residual Mean (SRM).
RQ3. What are some key requirements for using the method of paired comparisons to crowdsource a comprehensive relative risk of harm scale spanning hundreds of medications?

By doing this study, we learned about three sets of requirements that must be met to scale up paired comparison experiments with the CrowdSort web app to achieve large-scale crowdsourcing of expert knowledge by this method.

Access Requirements

The first set of requirements that must be met are access requirements. The first version of the CrowdSort web app surfaced several access issues which we overcame. However, CrowdSort is currently not accessible enough to be used for a large-scale experiment. Future versions of the CrowdSort web app need to enable paired comparison data collection on most web browsers and not only on the Google Chrome web browser that we used. Also, to scale up the use of CrowdSort, it needs to be upgraded to a responsive design that can be accessed on various devices including smart phones, tablets, laptops, and desktop computers. Furthermore, CrowdSort must be secured using a secure sockets layer (SSL) certificate. Also, CrowdSort must be made accessible to multiple simultaneous users.

Authorization Requirements

The second set of requirements relate to user authorization. For this study, we collected background information and made inquiries to ensure that every participant was a licensed pharmacist working in a role with patient care duties. Because we collected data remotely but synchronously during live video conference calls we could manually authorize pharmacists to participate. To expand the use of CrowdSort, a more robust and scalable method of validating user credentials is required. We are presently exploring options for automating the user authorization process.

Paired Comparison Task Division Requirements

The third set of requirements are intrinsic to the method of paired comparisons. To develop a fine-grained RRH scale for the majority commonly used medications requires 300 to 400 medications to be compared. Therefore, the number of head-to-head comparisons that must be made would range from 44,850 to 79,800. So many thousands of paired comparisons are far too many for individual pharmacists to complete them all. Instead, we are exploring ways to divide this sizeable comparison-making task up amongst a large number of pharmacist-participants, such as the cyclic paired comparison experiment designs described by David25.

Discussion

We began with a goal of demonstrating the feasibility of crowdsourcing a fine-grained scale for measuring the RRH of many medications. We confirmed that the method of paired comparisons can be used to acquire expert judgments about perceived medication worrisomeness remotely from a diverse group of practicing pharmacists. We are encouraged by the finding that our participants could complete 210 head-to-head medication comparisons in an
average of 16 minutes. This is evidence that, with effective software and incentives, the approach we have taken could scale up to support a large crowdsourcing program to estimate RRH for hundreds of medications.

In terms of granularity, our results for PMW (Figure 2) and for ITCW (Figure 3) demonstrate that it is possible to rank medications into multiple risk-informed PMW categories. By systematically probing pharmacists’ mental models of medication worrisomeness, we found statistically meaningful differentiation among five groups of medications and four groups of therapeutic drug classes. Perhaps more compelling is that these credible results came from engaging just 18 pharmacists. We hypothesize that more groups of medications will emerge on the PMW scale as additional paired comparison data enable more precise estimates. Besides, we did not seek to find definite lower or upper bounds for PMW. More work is needed to determine which medications are perceived to be least and most worrisome of all.

Our PMW scale has four different levels containing high-alert medications (Figure 2). Also, PMW does NOT consistently align with the distinction between prescription-only (i.e., Legend) versus over-the-counter (OTC) drugs. In our results, two Legend medications, amlodipine and amoxicillin, had the lowest PMW estimates. Meanwhile, three OTC medications – aspirin, ibuprofen (Motrin), and naproxen (Aleve) – had remarkably higher PMW estimates than amlodipine and amoxicillin. We did not ask the pharmacist-participants why they perceive two Legend medications to be less worrisome than three OTC drugs. It could be that, taken together, the relatively wide therapeutic windows of amoxicillin and amlodipine, the typically short duration of amoxicillin use, and the fact that amlodipine (Norvasc) is generally well tolerated, all led to these two medications being ranked least worrisome of the 21 in the study set.

It is clear that the 18 pharmacist-participants perceived anticoagulants to be the most worrisome of the study set, especially warfarin. The finding of high PMW and ITCW for anticoagulants may reflect nationwide safety problems with serious but avoidable bleeding events caused by warfarin and other anticoagulants. These findings suggest that a prescription for any oral anticoagulant may raise a mental warning flag in the minds of practicing pharmacists.

This study pertains to our prior published work examining how pharmacists currently develop medication-related risk scores and patient medication regimen complexity scores. Presently, many attempts to develop medication risk and complexity scores are localized at single organizations and involve assigning points to a small number of concerning medications, often with the most points going to those medications with the narrowest therapeutic windows. Few of these locally-devised medication-related scoring mechanisms have been validated. In contrast, our approach is to develop and validate a rigorous, comprehensive RRH scale for medications by combining the judgments of pharmacists from many organizations using the method of paired comparisons. In this manner, we seek to develop a robust, comprehensive, reliable, and valid RRH scale for assessing prescriptions and whole medication regimens.

More generally, this study highlights a potential role for the paired comparisons method in large-scale knowledge acquisition for learning health systems. To be fully successful, we believe learning health systems will need methods and mechanisms like CrowdSort to routinely and methodically source expertise from human practice on a grand scale. The new knowledge that results from crowdsourcing will need to be carefully tested and curated to make it useful.

**Limitations**

This study is limited in a variety of ways. First and foremost, a single experiment with 18 pharmacist-participants does not allow us to generalize our findings about the worrisomeness of medications in any meaningful way. We plan to do future work to expand, validate, and test the impact of these findings. Another limitation is that we only included 21 medications yet 300 to 400 medications need to be rank-ordered in terms of RRH to arrive at a useful measurement instrument. Finally, we were unable to include participant interviews in this study and so we cannot explain the logic of our pharmacist-participants with respect to the task of comparing the worrisomeness of medications head-to-head.

**Conclusion**

By engaging 18 practicing pharmacists from around the United States in a paired comparison experiment, we ascertained their individual and collective judgments about the perceived medication worrisomeness of 21 commonly prescribed medications. We performed data collection remotely using a web application developed for this study called CrowdSort. We found the pharmacists in our sample perceive the relative worrisomeness of medications in a systematic way giving rise to five distinct groupings of medications. This result shows it is feasible to group medications into multiple risk-related categories, transcending the binary categories of high-alert and NOT high-alert that we use in practice today. Further, we showed that some therapeutic categories are more worrisome than others in the minds of our pharmacist-participants. In preparation for future work, we identified several key requirements for scaling up this initial experiment. This study demonstrates the feasibility of developing a comprehensive fine-grained measure of the relative risk of harm by crowdsourcing expert knowledge from pharmacists.
Acknowledgements

We thank our pharmacist-participants. We are grateful to pharmacist Bruce Chaffee for motivating this work. We thank Rachel Kuo and Melanie Johnson for helping us test the Crowdsort web app. For guidance on our statistical approach, we thank Kerby Shedden and Corey Lester. For their general help in preparation for this experiment, we recognize and thank Betty Chaffee, Peter Boisvert, Charles P. Friedman, Nate Gittlen, and Brooke Raths.

References

13. Levinson DR. Hospital incident reporting systems do not capture most patient harm.2012, Jan, AHRQ.
The Use of Inter-terminology Maps for the Creation and Maintenance of Value Sets

Kin Wah Fung¹, MD, MS, MA; Julia Xu¹, MD, PhD; Sigfried Gold¹, MFA, MA

¹National Library of Medicine, National Institutes of Health, Bethesda, MD

tkwfung@nlm.nih.gov|xujc@mail.nlm.nih.gov|sigfried@sigfried.org

Abstract

Value sets are essential in activities such as electronic clinical quality measures (eCQM) and patient cohort definition. Creation and maintenance of value sets is labor intensive and error prone. Our method aims to use existing inter-terminology maps to improve the quality of value sets that are defined in more than one terminology. For 197 eCQM value sets defined in SNOMED CT plus ICD-9-CM and/or ICD-10-CM, the map-generated codes showed good overlap with the value set codes. Manual review showed that some new codes identified by mapping should probably be included in the value sets. This could potentially augment the ICD-9-CM codes by 45% (1.5 codes), ICD-10-CM codes by 25% (1.8 codes) and SNOMED CT codes by up to 42% (4.8 codes) per value set on average. The mapping between SNOMED CT and ICD-10-PCS did not perform as well because of the granularity discrepancy in the map.

Introduction

A value set is typically a list of codes taken from a biomedical terminology that collectively defines the scope of a clinical concept. Value sets can be used to identify patient cohorts, measurement criteria for clinical quality improvement, allowable values for a data element (in a survey instrument or the electronic health record), among other uses. One important function of value sets is to ensure that health information from disparate sources are interoperable, whether to support collective data analytics or continuity of patient care. Towards this goal, value set codes are often derived from commonly used clinical terminology standards such as SNOMED CT, ICD code sets (ICD-9-CM, ICD-10-CM and ICD-10-PCS), LOINC and RxNorm.

Value sets play a crucial but often unappreciated and unnoticed role in the capture, use, and analysis of clinical information. Of all the professionals who interact with and rely on value sets in their work with health records, a small fraction is likely to know what a value set is, even in the informatics community, which has a deep, foundational appreciation for the role of standardized terminologies in the complex ecosystem of medical care and health sciences.

Creation of value sets is a labor-intensive process that usually involves subject matter experts working with terminologists to identify codes to include (and exclude) for a particular use case. Maintenance of value sets is essential to keep them up-to-date, since terminologies do evolve over time. The burden of curation is even bigger for value sets that are defined by more than one terminology because the same clinical information can be encoded by different code systems e.g., SNOMED CT and ICD-10-CM for diagnosis. Previous studies have investigated automated methods to assist the authoring and quality assurance of value sets, making use of information such as the hierarchical structure of terminologies and the semantic types of the codes.¹⁻³ This study describes a novel approach of using existing inter-terminology maps to improve the quality of multi-terminology value sets. In an earlier study, we have shown that published maps between ICD-9-CM and ICD-10-CM can help in the translation of codes in value sets.⁴

High quality and up-to-date value sets will generally increase their re-use and utility.⁵ This paper concentrates on one domain where value set reuse is assured by mandate and incentive: as parts of eCQM (electronic clinical quality measures).⁶ The Centers for Medicare & Medicaid Services (CMS) use eCQM in a variety of quality reporting and value-based purchasing programs. The main advantage of electronic quality measurement is that the burden of manual chart pulling, data extraction and reporting can be much reduced. In addition, there is reduced lag time from documentation to reporting and potential access to real time data for quality measurement.

The National Library of Medicine (NLM) in collaboration with the Office of the National Coordinator for Health Information Technology (ONC) and CMS runs the Value Set Authority Center (VSAC), a repository, website, and set
of application programming interfaces (APIs) for public value sets created by external parties. The VSAC does not create value set content but provides tools for value set authors to create and maintain their value sets. Through the VSAC, users can download all official, up-to-date versions of eCQM value sets. VSAC also hosts value sets from other sources, such as the HL7 C-CDA (Consolidated Clinical Document Architecture) and CMS Core Clinical Data Elements and Hybrid Measures.

**Methods**

We downloaded all the eCQM value sets from the VSAC website. Among the eCQM value sets, we identified those that were defined using more than one terminology. One of the terminologies had to be SNOMED CT, and the other could be one or more of ICD-9-CM, ICD-10-CM and ICD-10-PCS. ICD-9-CM included both diagnosis and procedure codes. We excluded the ICD-9-CM procedure codes, which were used in only a small number (2 out of 140) of value sets with ICD-9-CM codes. In addition to the lists of codes, the download file also contained metadata about the value set, including the value set OID, value set name, version and purpose. The purpose described the clinical focus, data element scope, inclusion and exclusion criteria. However, the purpose was not always filled in.

Figure 1. Overall schema of study

---

**Generation of the reference inter-terminology maps**

1. **SNOMED CT and ICD-9-CM map**

   To generate the map between SNOMED CT and ICD-9-CM, we combined two sources of mappings. The first was the ICD-9-CM Diagnostic Codes to SNOMED CT Map published by NLM. The purpose of this map is to facilitate translation of ICD-9-CM codes to SNOMED CT. Even though the U.S. moved from ICD-9-CM to ICD-10-CM in 2015, many existing electronic health record systems still contain clinical information encoded in ICD-9-CM. To facilitate migration to SNOMED CT as the primary clinical terminology for patient health problems (diseases and conditions), it is desirable that the legacy ICD-9-CM data be translated to SNOMED CT. This map is updated yearly to synchronize with the latest version of SNOMED CT. This map is divided into two parts: the one-to-one and one-to-many maps. From both parts, we extracted pairs of SNOMED CT and ICD-9-CM codes into our map. The second source of mappings was the SNOMED CT to ICD-9-CM map published by SNOMED International. The original goal of this map was to facilitate the generation of ICD-9-CM codes from SNOMED CT-encoded clinical information. Since most countries using ICD-9-CM have moved to ICD-10-CM, SNOMED International stopped the maintenance of this map in 2016. We retrieved the last version of this map through the MRMAP table in the UMLS (2016AA version). We harvested pairs of SNOMED CT and ICD-9-CM codes from this map, excluding SNOMED CT concepts that have become inactive since the last update of the map. The combined set of code pairs formed our **SCT-I9 Map**.
2. SNOMED CT and ICD-10-CM map

We used the SNOMED CT to ICD-10-CM map published by NLM. This is a rule-based map that maps all SNOMED CT concept from three hierarchies (Clinical finding, Event and Situation with explicit context) to ICD-10-CM. The map is rule-based to cater for the ICD-10-CM coding rules, which sometimes stipulate that the same disease be coded differently according to patient age, gender and co-morbidities. For example, the SNOMED CT concept Failure to gain weight (36440009) may be coded to ICD-10-CM codes Adult failure to thrive (R62.7), Failure to thrive in newborn (P92.6) or Failure to thrive (child) (R62.51) depending on age. In the NLM map, there is always a default map target that is used when no additional information is available. For our study, we harvested all pairs of SNOMED CT and their default ICD-10-CM map targets to create our SCT-I10 Map.

3. SNOMED CT and ICD-10-PCS map

For mapping between SNOMED CT and ICD-10-PCS, we used the mappings contained in the vocabulary resource of the OHDSI Consortium (Observational Health Data Sciences and Informatics). The purpose of the OHDSI vocabulary is to enable transparent and consistent content across disparate observational databases to support efficient and reproducible research. The mappings are expressed as ‘is a’ relationships between an ICD-10-PCS code and a SNOMED CT concept. For example, the ICD-10-PCS code Bypass Cerebral Ventricule to Intestine with Synthetic Substitute, Open Approach (00160J5) is a Ventriculostomy (63933000) in SNOMED CT. We harvested all the pairs of SNOMED CT and ICD-10-PCS codes linked by ‘is a’ to form our SCT-PCS Map.

Identification of map-generated code lists for each value set using the reference maps

For each value set, we used the reference maps to find mappings for the original value set codes in each terminology. We called the codes found by mapping the “map-generated codes”. We used the reference maps in either direction. For example, if the value set contained SNOMED CT, ICD-9-CM and ICD-10-CM codes, we would generate the following four lists of map-generated codes

- **I9-from-SCT codes** - ICD-9-CM codes identified based on the original SNOMED CT codes in the value set mapped through the SCT-I9 Map
- **I10-from-SCT codes** - ICD-10-CM codes identified based on the original SNOMED CT codes in the value set mapped through the SCT-I10 Map
- **SCT-from-I9 codes** – SNOMED CT codes identified based on the original ICD-9-CM codes in the value set mapped through the SCT-I9 Map
- **SCT-from-I10 codes** – SNOMED CT codes identified based on the original ICD-10-CM codes in the value set and mapped through the SCT-I10 Map

Evaluation of the map-generated codes

For each value set, we evaluated the map-generated codes in two aspects. First, we wanted to know how closely the map-generated codes resembled the original value set codes. Second, we wanted to know whether the map-generated codes could help to identify codes that should have been included in the value set. The evaluation was done as follows:

1. **Jaccard similarity coefficient**

   We used the Jaccard coefficient as an indicator of the similarity between the original value set codes and the map-generated codes for each terminology of each value set. The Jaccard coefficient was defined as the number of codes common to both the original and map-generated code lists divided by the number of codes in the union of the two lists. The Jaccard coefficient generally reflects the degree of overlap between two lists of items in which the number of items can be different. As an example, for a value set with SNOMED CT, ICD-9-CM and ICD-10-CM codes, we would generate four Jaccard scores between:
   a. original ICD-9-CM and I9-from-SCT codes
   b. original ICD-10-CM and I10-from-SCT codes
   c. original SNOMED CT and SCT-from-I9 codes
   d. original SNOMED CT and SCT-from-I10 codes
2. Manual review
   We performed manual review of all the codes found by mapping that were not present among the original codes in a value set. For practical purposes, we limited our review to cases in which the total number of codes (both map-generated and original) for the pair of terminologies in focus was less than 100. The reviewer judged whether the new code found by the map was:
   a. Exact match – the map-generated code matched exactly the meaning of one of the original codes in the value set. The map-generated code should probably be included in the value set.
   b. Approximate match, candidate – the map-generated code was similar in meaning to at least one original code in the value set, and should be considered as a candidate for addition to the value set. We judged the suitability of inclusion based on the stated purpose of the value set and the inclusion criteria, if explicitly stated. Otherwise, we would rely on the name of the value set and other codes in the value set to make our judgment.
   c. Approximate match, not a candidate – the map-generated code was similar in meaning to at least one original code in the value set, but was not considered a candidate for addition due to the inclusion and exclusion criteria, if explicitly stated.
   d. Not a match – the map-generated code did not match any of the original value set codes.

   In interpreting the review results, we assumed that all codes in a terminology that were within the scope of the value set (as judged by the name, accompanying documentation and the original value set codes) should be included in the value set. We assumed that there was no exclusion for certain special kind of codes (e.g., unspecified codes) unless explicitly stated. We assumed that the code list in one terminology in a value set could be used independently from other code lists in a different terminology. Finally, we also assumed that map-generated codes that did not fall within the scope of the value set and did not match any of the original value set codes should not be included in the value set.

Results

The value sets

We downloaded 715 eCQM value sets (version September 2018) from VSAC of which 221 contained SNOMED CT and one or more of the other target terminologies. Table 1 shows the combination of terminologies in these value sets.

Table 1. Distribution of terminologies among the value sets

<table>
<thead>
<tr>
<th>SNOMED CT</th>
<th>ICD-9-CM</th>
<th>ICD-10-CM</th>
<th>ICD-10-PCS</th>
<th>Number of value sets with this combination of terminologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>✓</td>
<td>✓*</td>
<td>✓</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Number of value sets with this terminology: 221

*2 value sets using ICD-9-CM procedure codes excluded from study

Among the 221 value sets, only 83 (38%) had explicitly stated purpose, scope, inclusion and exclusion criteria.

In the following section, we present our findings according to the pair of terminologies covered by the reference maps that we created.
1. SNOMED CT and ICD-9-CM map

The SCT-I9 Map we created could provide mapping to 89,043 SNOMED CT concepts, which represented 72% of concepts in the three relevant hierarchies (Clinical finding, Event and Situation with special context). On the ICD-9-CM side, it covered 13,014 (89%) diagnostic ICD-9-CM codes.

Table 2. Results of mapping between SNOMED CT and ICD-9-CM

<table>
<thead>
<tr>
<th>Direction in which the SCT-I9 Map is used</th>
<th>Number of value sets:</th>
<th>Value sets with map-generated codes (%)</th>
<th>Number of codes:</th>
<th>Map-generated codes found in value sets (%)</th>
<th>Jaccard score: macro-average (median)</th>
<th>Manual review:</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9-CM from SNOMED CT</td>
<td>138</td>
<td>136 (99%)</td>
<td>3,162 ICD-9-CM codes</td>
<td>2,287 (62%)</td>
<td>0.5 (0.5)</td>
<td>87</td>
</tr>
<tr>
<td>SNOMED CT from ICD-9-CM</td>
<td>138</td>
<td>138 (100%)</td>
<td>9,964 SNOMED CT codes</td>
<td>8,203 (53%)</td>
<td>0.38 (0.35)</td>
<td>89</td>
</tr>
</tbody>
</table>

Table 2 summarizes the results of using the SCT-I9 Map to generate candidate value set codes. The map was able to suggest some codes in almost all the value sets in either direction of mapping. For the map-generated codes, 62% of the ICD-9-CM codes and 53% of SNOMED CT codes already existed in the value set. For each value set, we measured the similarity between the original codes in the value set and the map-generated codes for the same terminology by calculating the Jaccard coefficient. The average of the Jaccard coefficients for ICD-9-CM and SNOMED CT was 0.5 and 0.38 respectively. We manually reviewed all the map-generated codes that were not among the original value set codes, limiting to value sets with 100 codes or less. We reviewed a total of 305 ICD-9-CM and 789 SNOMED CT codes from 89 value sets.

For the map-generated ICD-9-CM codes, 33% of the codes not present in the original value set were found to be exact matches to another code in the value set. For example, in the Optic Neuritis value set, the ICD-9-CM code *Meningococcal optic neuritis (036.81)* was found by mapping but not in the original value set. However, the SNOMED CT code *Meningococcal optic neuritis (73431005)* was in the original value set. A further 16% of the map-generated ICD-9-CM codes were approximate matches and could be considered candidates for addition. For example, in the Uveitis value set, the ICD-9-CM code *Syphilitic uveitis, unspecified (091.50)* was found by mapping but not in the original value set. This code should probably be added to the value set since it also contained the SNOMED CT code *Uveitis due to secondary syphilis (186854007)*. Overall, 50% of the reviewed ICD-9-CM codes did not match any of the value set codes.

For the map-generated SNOMED CT codes, 4% of the codes not present in the original value set were found to be exact matches. For example, in the Hypotension value set, the SNOMED CT code *Orthostatic hypotension (28651003)* was found by mapping but not in the original value set. However, the ICD-9-CM code *Orthostatic hypotension (458.0)*
was in the original value set. Almost half (45%) of the map-generated SNOMED CT codes were approximate matches and candidates for addition. For example, in the Bradycardia value set, the SNOMED CT code Symptomatic sinus bradycardia (444605001) was found by mapping but missing from the value set. This code should be considered a candidate to be added to the value set since it also contained the SNOMED CT code Severe sinus bradycardia (49044005). A small number of codes (2%) were approximate matches but not considered candidates for addition. This number could be artificially low because only a third of all value sets had explicitly stated purpose, inclusion and exclusion criteria, which were needed to determine if a code belonged to this category. For example, in the Lupus value set, the SNOMED CT code Systemic lupus erythematosus in remission (698694005) was an approximate match but not a candidate because the exclusion criteria explicitly excluded “systemic lupus erythematosus in childhood or in remission”. Overall, 48% of the reviewed SNOMED CT codes did not match any of the value set codes.

The unmatched codes generally belonged to two types. The first type was mapped-generated codes based on a part of a composite term that was not relevant to the value set. For example, the Lupus value set contained the SNOMED CT code Pericarditis co-occurrent and due to systemic lupus erythematosus (25380002), which was mapped to the ICD-9-CM code Other acute pericarditis (420.99) that was outside the scope of the value set. The second type was related to the “Not Elsewhere Classified” (residual or catch-all) codes in ICD-9-CM. For example, the Bradycardia value set contained the ICD-9-CM code Other specified cardiac dysrhythmias (427.89) which was mapped to many SNOMED CT codes, including Sinus tachycardia (11092001) that were not within the scope of the value set.

2. SNOMED CT and ICD-10-CM map

The SCT-I10 Map we created could provide mapping to 110,184 SNOMED CT concepts, which represented 89% of concepts in the three relevant hierarchies (Clinical finding, Event and Situation with special context). On the ICD-10-CM side, it covered 30,576 (43%) ICD-10-CM codes.

Table 3. Results of mapping between SNOMED CT and ICD-10-CM

<table>
<thead>
<tr>
<th>Number of value sets:</th>
<th>ICD-10-CM from SNOMED CT</th>
<th>SNOMED CT from ICD-10-CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>197</td>
<td>193 (98%)</td>
<td>176 (89%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of codes:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original codes in value sets</td>
</tr>
<tr>
<td>Map-generated codes</td>
</tr>
<tr>
<td>Map-generated codes found in value sets (%)</td>
</tr>
</tbody>
</table>

| Jaccard score: macro-average (median) | 0.3 (0.2) |
| Manual review: |
| Value sets reviewed | 112 |
| Original codes in value sets | 806 ICD-10-CM codes |
| Map-generated codes reviewed | 547 ICD-10-CM codes |
| Exact match (%) | 136 (25%) |
| Approximate match, candidate (%) | 127 (23%) |
| Approximate match, not candidate (%) | 0 (0%) |
| Not a match (%) | 284 (52%) |

Table 3 summarizes the results of using the SCT-I10 Map to generate candidate value set codes. The map was able to suggest some ICD-10-CM and SNOMED CT codes in 98% and 89% of the value sets respectively. For the map-
generated codes, 59% of the ICD-10-CM codes and 37% of SNOMED CT codes already existed in the value set. The macro-average of the Jaccard coefficients for ICD-10-CM and SNOMED CT was 0.3 and 0.32 respectively.

We reviewed 547 ICD-10-CM and 1,124 SNOMED CT codes (from 112 and 96 value sets) that were not present in the original value sets. For the map-generated ICD-10-CM codes, 25% of the codes were found to be exact matches to another code in the value set. For example, in the *Chlamydia* value set, the ICD-10-CM code *Chlamydial peritonitis (A74.81)* was found by mapping but not in the value set. However, the SNOMED CT code *Chlamydial peritonitis (197172005)* was in the original value set. A further 23% of the map-generated ICD-10-CM codes were approximate matches and could be considered candidates for addition. For example, in the *Breastfeeding* value set, the ICD-10-CM code *Hypogalactia (O92.4)* was found by mapping but not in the value set. This code should be considered for addition to the value set since it also contained the ICD-10-CM code *Suppressed lactation (O92.5)*. Overall, 52% of the reviewed ICD-10-CM codes did not match any of the value set codes.

For the map-generated SNOMED CT codes, 5% of the codes not present among the original value set codes were found to be exact matches. For example, in the *Rubella* value set, the SNOMED CT code *Rubella meningitis (1092351000119107)* was found by mapping but not in the value set. However, the ICD-10-CM code *Rubella meningitis (B06.02)* was in the value set. Almost three-quarter (72%) of the map-generated SNOMED CT codes were approximate matches and candidates for addition. For example, in the *Proteinuria* value set, the SNOMED CT code *Microalbuminuria due to type 1 diabetes mellitus (18521000119106)* was found by mapping but missing from the value set. This code should be considered a candidate to be added to the value set since it also contained the SNOMED CT code *Microalbuminuria (312975006)*. A small number of codes (0.4%) were approximate matches but not candidates for addition. For example, in the *Chronic Malnutrition* value set, the SNOMED CT code *Progressive encephalopathy with severe infantile anorexia (715794009)* was an approximate match but not considered a candidate because of the exclusion criteria “all diagnosis codes for feeding disorders of infancy or childhood”. Overall, 23% of the reviewed SNOMED CT codes did not match any of the value set codes.

The unmatched codes could generally be attributed to the same reasons as for the SCT-I9 Map (see above).

3. **SNOMED CT and ICD-10-PCS map**

The SCT-PCS Map we created could provide mapping to 4,563 SNOMED CT concepts, which represented 8% of concepts in the Procedure hierarchy. On the ICD-10-PCS side, it covered 40,378 (51%) ICD-10-PCS codes.

Table 4 summarizes the results of using the SCT-PCS Map to generate candidate value set codes. The map was able to suggest some ICD-10-PCS codes in 83% of the value sets, and some SNOMED CT codes in all value sets. For the map-generated codes, 47% of the ICD-10-PCS codes and 34% of SNOMED CT codes already existed in the value set. The macro-average of the Jaccard coefficients for ICD-10-PCS and SNOMED CT was 0.38 and 0.08 respectively.

We reviewed a total of 99 ICD-10-PCS and 48 SNOMED CT codes (from 8 and 12 value sets) that were not present in the original value sets. For the map-generated ICD-10-PCS codes, 15% were approximate matches and could be considered candidates for addition. For example, in the *Hip Fracture Surgery* value set, the ICD-10-PCS code *Repair Right Hip Joint, Percutaneous Endoscopic Approach (0SQ94ZZ)* was found by mapping but not in the original value set. This code should be considered for addition to the value set since it also contained the ICD-10-PCS code *Repair Right Hip Joint, Open Approach (0SQ90ZZ)*. The remaining 85% of the reviewed ICD-10-PCS codes did not match any of the value set codes.

For the map-generated SNOMED CT codes, 19% of the codes not present in the original value set were found to be approximate matches and candidates for addition. For example, in the *BH Outpatient Psychotherapy* value set, the SNOMED CT code *Individual psychotherapy (18512000)* was found by mapping but missing from the value set. This code could be considered a candidate to be added to the value set which also contained the SNOMED CT code *Group psychotherapy (76168009)*. The remaining 81% of the reviewed SNOMED CT codes did not match any of the value set codes.
Table 4. Results of mapping between SNOMED CT and ICD-10-PCS

<table>
<thead>
<tr>
<th></th>
<th>Direction in which the SCT-PCS Map is used</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICD-10-PCS from SNOMED CT</td>
</tr>
<tr>
<td><strong>Number of value sets:</strong></td>
<td>23</td>
</tr>
<tr>
<td>Value sets with map-generated codes (%)</td>
<td>19 (83%)</td>
</tr>
<tr>
<td><strong>Number of codes:</strong></td>
<td></td>
</tr>
<tr>
<td>Original codes in value sets</td>
<td>4,507 ICD-10-PCS codes</td>
</tr>
<tr>
<td>Map-generated codes</td>
<td>3,337 ICD-10-PCS codes</td>
</tr>
<tr>
<td>Map-generated codes found in value sets (%)</td>
<td>1,585 (47%)</td>
</tr>
<tr>
<td>Jaccard score: macro-average (median)</td>
<td>0.38 (0.27)</td>
</tr>
<tr>
<td><strong>Manual review:</strong></td>
<td></td>
</tr>
<tr>
<td>Value sets reviewed</td>
<td>8</td>
</tr>
<tr>
<td>Original codes in value sets</td>
<td>57 ICD-10-PCS codes</td>
</tr>
<tr>
<td>Map-generated codes reviewed</td>
<td>99 ICD-10-PCS codes</td>
</tr>
<tr>
<td>Approximate match, candidate (%)</td>
<td>15 (15%)</td>
</tr>
<tr>
<td>Not a match (%)</td>
<td>84 (85%)</td>
</tr>
</tbody>
</table>

Most of the unmatched codes were related to the fact that the majority of the SNOMED CT codes in the SCT_PCS Map were broader than any of the value set codes (see Discussion).

**Discussion**

Value sets are useful resources and help to ensure the validity and comparability of data collected for various purposes. However, the creation and maintenance of value sets is a non-trivial process. Terminologies are constantly updated to reflect advance in biomedical science and changes in coding rules. While it is relatively straightforward to identify codes that have become obsolete when a terminology is updated, it is more difficult to identify new codes that should be included in a value set. In medical classifications like ICD-9-CM or ICD-10-CM, due to the requirement of a single (as opposed to a poly-) hierarchy, new codes related to the same condition are sometimes added to a different branch from the existing codes (e.g., in ICD-9-CM, hypertension in pregnancy is put under the chapter Complications of pregnancy, childbirth and puerperium), making them more liable to be missed by the value set editors. If value sets are not properly maintained, their quality will deteriorate over time, with deleterious effects on the validity of the data generated using these value sets. The impact can range from inaccurate calculation of payments to providers due to errors in electronic clinical quality reporting, to missing patients from a research cohort due to an omitted code.

Our study shows that existing inter-terminology maps can be useful adjuncts for the creation and maintenance of value sets that involve more than one terminology. Take the example of mapping between SNOMED CT and ICD-9-CM. There are 138 eCQM value sets that are defined by both of these terminologies. The overlap between the original value set codes and map-generated codes is considerable judging from the Jaccard scores (0.5 and 0.38 for ICD-9-CM and SNOMED CT respectively). This lends support to the general validity of our method that a significant portion of the original value set codes can be found by inter-terminology mapping. More importantly, mapping also uncovers a significant number of codes that are potentially missed by the value set curators. Based on our review, the exact matches are almost certainly missed codes that should be added. We think that a considerable portion of the approximate matches should probably be added as well. If we assume that all the exact matches and half of the approximate match candidate codes are indeed added to the value sets, the SCT-19 Map would augment the 87 value sets by a total of 127 ICD-9-CM codes (1.5 codes per value set). This represents a 45% increase to the 284 original ICD-9-CM codes in these value sets. (Table 2) By similar assumptions, the SNOMED CT codes in these value sets would be augmented by 21% (2.4 codes per value set) on average. Similarly, the SCT-110 Map will augment the ICD-
10-CM codes by 25% (1.8 codes per value set) and SNOMED CT codes by 42% (4.8 codes per value set) on average. (Table 3)

The results for the SCT-PCS Map are less satisfactory. The overlap between the map-generated codes and the value set codes is on the low side, especially in the ICD-10-PCS to SNOMED CT direction. Manual review did not find any exact matches. This is not surprising considering the differences between SNOMED CT and ICD-10-PCS. ICD-10-PCS codes are generally more granular and contain information (e.g., surgical approach, use of devices and procedure intent) that is not captured in SNOMED CT concepts. Therefore, it is unusual to find an exact match between the two terminologies. The main purpose of the mappings in the OHDSI vocabulary is to support data aggregation and analytics. Many of the SNOMED CT concepts in the map are very general (e.g., Procedure on skin (118718002)), whereas value sets tend to use more specific SNOMED CT concepts. Moreover, the mappings in the OHDSI vocabulary are generated algorithmically with limited human review, unlike the other mapping resources used in this study that are all manually validated. It is possible that the performance of the SCT-PCS Map can be improved by propagating the mappings of the broader SNOMED CT concept to their descendants. We shall explore this option in future.

Use of inter-terminology maps in the creation and maintenance of value sets is likely to improve their quality (e.g., reducing missing codes) and reduce the curation time and effort. All the maps used in this study are in the public domain and regularly updated by the owners (except for the SNOMED CT to ICD-9-CM map from SNOMED International). To be useful, the codes suggested by the maps should have high accuracy (i.e. high signal-to-noise ratio). In our study, the proportion of map-generated new ICD-9-CM, ICD-10-CM and SNOMED CT codes that are either exact or approximate candidate matches are 49%, 48%, and from 49% (mapped from ICD-9-CM) to 77% (mapped from ICD-10-CM) respectively. (Tables 2 and 3) The suggestions based on the SNOMED CT to ICD-10-PCS map are not as useful. In actual implementation, how to present the map-suggested codes to the value set editors in a dynamic and efficient way is an important consideration. We hope to explore this issue in our future research. A potential refinement of our method is to use multiple routes of mapping to hone in on the most promising codes. For example, if a value set contains SNOMED CT, ICD-9-CM and ICD-10-CM codes (the most common type in our study), the SNOMED CT codes generated from the ICD-9-CM and ICD-10-CM maps can be used for mutual validation.

The VSAC has become the de facto home for dissemination of value sets for various purposes. It is also an authoring platform for value sets. The recent addition of the ability to create intensional value sets is likely to reduce the burden of value set curation. Instead of enumerating the codes in a value set (extensional value set), an intensional value set specifies the rule for inclusion of codes. An example will be using the expression ‘377.3-’ instead of a list of ICD-9-CM codes (‘377.30’, ‘377.31’, ‘377.32’, ‘377.33’, ‘377.34’, ‘377.39’) to represent the value set of Optic neuritis. When a terminology is updated, it is expected that the intensional definition will pick up new codes automatically, provided that they are subsumed by the same parent codes. However, this is more likely to be true in a multi-hierarchical terminology like SNOMED CT. How useful it is for single-hierarchy terminologies like ICD-9-CM and ICD-10-CM remains to be seen. Another observation in our study is that only 38% of the eCQM value sets have explicitly stated purpose and inclusion/exclusion criteria. This kind of information is essential for users to understand the scope of a value set and determine whether it fits their use case. Without this information, the users can only guess by the name of the value set and the codes it contains. For reliable re-use of value sets, it is essential that value set authors provide the necessary value set metadata and documentation.

We recognize the following limitations in our study. The study only focused on the eCQM value sets that are defined in more than one terminology. The results may not be generalizable to other value sets. We only studied maps between four terminologies, albeit that they are among the most commonly used in value sets. We found significant performance difference among different terminologies. Whether this method is applicable to other terminologies remains to be seen. All manual review of the map-generated codes was done by one author (JX) and not independently validated. We only reviewed value sets with 100 codes or less.
Conclusion

Inter-terminology maps can potentially be used to assist human curators in the creation and maintenance of value sets that are defined by multiple terminologies. The map-generated codes showed good overlap with the original value set codes, lending support to the validity of the method. For value sets using SNOMED CT in combination with ICD-9-CM and/or ICD-10-CM, the map-generated codes could potentially augment the ICD-9-CM codes by 45% (1.5 codes per value set), ICD-10-CM codes by 25% (1.8 codes per value set) and SNOMED CT codes by up to 42% (4.8 codes per value set) on average. The map between SNOMED CT and ICD-10-PCS did not work as well, probably due to the big difference in granularity between the terminologies and the inexact nature of the source mappings.

Acknowledgement

This work was supported in part by the Intramural Research Program of the National Institutes of Health and the National Library of Medicine.

References


Disseminating Strengths-Oriented Best Practices in Diabetes Care
Utilizing a Standardized Language within a Global Community

Grace Gao, PhD, DNP, RN-BC1, Madeleine Kerr, PhD, RN2, Sarah Beman, PhD, RN, PHN, CNE1, Candice Bruhjell, PhD, RN, FCN1, Joyce Rudenick, DNP, MAHS, RN-BC1, Onkar Singh, PhD, RN1, Mehrdad Rafiei, PhD1, Karen A. Monsen, PhD, RN, FAMIA, FAAN 2

1Saint Catherine University, St. Paul, Minnesota, USA; 2University of Minnesota, Minneapolis, Minnesota, USA

Abstract

Diabetes is a manageable chronic condition that contributes significantly to the global health burden of diseases and mandates a global collective effort to create an effective solution. This paper describes a community diabetes care pathway built upon a Strengths-Oriented Global Health Informatics Framework and an interdisciplinary standardized terminology, the Omaha System, along with a related translational process to disseminate best practices in diabetes care in China. This project demonstrates a novel strengths-oriented collaborative approach to disseminate best practices of diabetes management in global health communities and offers a potential to bring person-centered coordinated care to multi-levels of engagement that generate actionable and measurable results. Such collaboration opens a continued dialogue in the discourse for constructing global health informatics principles and practice to reduce the burden of diseases around the world.

Introduction

As one of the four priority noncommunicable diseases (NCDs), diabetes contributes significantly to the global health burden of diseases, reflective of a surge in associated risk factors such as obesity in the general population globally.1 Diabetes was a direct cause of death for an estimated 1.6 million people in 2016.2 As a major global health issue, diabetes in China is now becoming the world’s biggest epidemic.3 With the estimated 11 percent population suffering diabetes,4 China has the most diabetes cases worldwide. In company with its strong economic growth, pressing needs have emerged to promote diabetes care in China. To address this increasing health challenge and reduce the burden of diabetes around the globe, commitments among nations have been sought for the development of a collective consensus as well as investment in creating affordable, cost-effective interventions based on the best available science.1

Global health informatics can be used as an effective vehicle to promote what is defined by global health as improving health and health equity for all people worldwide5 via dissemination of best global health informatics practices and principles and support of an exchange forum of global health experiences and expertise.6 This new global social space not only strengthens connections around the world but also cultivates locality and regionality through globalization.7 Multi-partnerships through global efforts have been implemented in the past to improve diabetic care in China based on a Share-Care Community Model.8 Community-based care has been given a global focus on the initiation of best community practices through the employment of the standardized language, the Omaha System, by the Omaha System Community of Practice.9 This paper describes the creation of an open source diabetes care pathway for community care based on a Strengths-Oriented Global Health Informatics Framework and an interdisciplinary standardized terminology, the Omaha System; and further describes a related translational process to disseminate best practices in diabetes care in China, a country with the most populous cases of diabetes in the world. This project is a prototype of an interdisciplinary strengths-based diabetes care plan to meet the increasing need for reducing the burden of NCDs in a global health setting.

Strengths-Oriented Global Health Informatics Framework and the Omaha System

Health strengths or health assets are defined as any protective and promoting factor or resource that helps individuals, communities, and populations to maintain and sustain health and wellbeing, and reduce health inequalities.10-12 A strengths-based approach aims to highlight evidence of effectiveness to address problems or eliminate deficits with emphasis on positive attributes.12 Such approaches strive to identify protective factors and resources to mobilize solutions that support and promote the health and wellbeing of both individuals and communities.12,13 A limited number of robust evaluations of assets-based actions exists which challenges
professionals to rethink strategies for disease management and health promotion. Building a strengths or assets-based framework to guide care interventions in the management of diabetes for global health communities offers a strong asset to turn interventions and actions into evaluations that generate data and information into measurable knowledge and practice outcomes.

Health care services have traditionally been designed to fix actual or potential health problems of individuals or populations by providing interventions to solve, alleviate, or prevent those problems, needs, and deficiencies, with little focus on enhancing the individuals’ strengths and capabilities or incorporating an individual’s experiences, preferences, perspectives and knowledge. Such an approach places individuals as passive recipients of services rather than active participants in their own care. Leveraging the positive capabilities, strengths, and resources of individuals may allow them to identify problems and prompt solutions for health and wellbeing that promote resilience and improved health outcomes. Health promoting or protecting assets can be harnessed on individual, community, and organizational levels across the domains of health determinants that include individual characteristics, social circumstances, environmental conditions of living and work, behavioral choices related to health and health services, and equitable and sustainable social and economic development.

A prototype of a Strengths-Oriented Global Health Informatics Framework was developed on three levels: individuals, communities/populations, communities/organizations. At the level of individuals, health promoting or protecting assets span across the domains of health determinants including personal, individual characteristics, and health behaviors. At the level of communities and populations, health promoting or protecting assets originate from social circumstances, environmental conditions of living and working, behavioral choices related to health and health services. At the level of communities and organizations, health promoting or protecting assets stem from equitable and sustainable social and economic development.

This Strengths-Oriented Global Health Informatics Framework was then mapped to and represented by the Omaha System Domains (Figure 1). The Omaha System is an interdisciplinary standardized terminology intended to capture health and health care by taxonomically describing all of health as 42 defined Problem concepts within four holistic domains: Environmental, Psychosocial, Physiological, and Health-related Behaviors. Multidisciplinary clinicians use Omaha System point-of-care software in the United States and other countries created by a growing number of computer software vendors who offer Omaha System clinical information systems. This standardized vocabulary is used predominantly in the community and public health settings to document care assessments, interventions, and outcomes.

The Omaha System consists of three classification schemata including the Problem Classification Scheme, the Intervention Scheme, and the Problem Rating Scale for Outcomes. The Problem Classification Scheme has 4 standardized domains that branch into a total of 42 standardized problem concepts which describe assessments for problem-specific strengths and signs/symptoms. The Intervention Scheme is related to the Problem Classification Scheme and has 3 hierarchical levels including 4 standardized categories in action terms and 75 standardized targets followed by customizable care descriptions to document care plans and services for each identified problem concept. The Problem Rating Scale for Outcomes measures the problem concepts using three 5-point Likert-type scales (1 =
lowest to 5 = highest) related to Knowledge, Behavior, and Status (KBS) outcomes. Each of the KBS rating scales presents a continuous framework for evaluating a distinct rating of a problem concept and an intervention outcome over time. It is feasible to use the Omaha System to document both problems and strengths on individual and community levels. The reliability and validity of the Omaha System components have been tested and established in multiple settings in research.

Care pathways encoded with the Omaha System have been integrated into electronic health records and mobile applications. Similarly, the current strengths-oriented diabetes care pathway using the Omaha System can be easily integrated into a web-based or mobile platform to support diabetes care management on the go. Using mobile technology is critical as it has dominated the everyday life of citizens in China. Leveraging the use of technology and data offers the potential in discovery of new paradigms in tracking and managing disease, defining best practice that integrates a patient-centered perspective, addressing quality gaps, and demonstrating improvement in health outcomes.

The Omaha System Strengths-Oriented Diabetes Care Pathway

A Strengths-Oriented Diabetes Care Pathway (Figure 2) was created for the management of diabetes in the global health setting by referencing to evidence-based clinical guidelines developed for standardized diabetes care in the U.S, expert opinions, and literature. Based on the Omaha System domains, the pathway is mapped to the three levels of individuals, communities/populations, and communities/organizations for the Strengths-Oriented Global Health Informatics Framework. This Strengths-Oriented Diabetes Care Pathway identifies both problems, the deficits, and strengths, the health assets and positive attributes using relevant Omaha System problem concepts corresponding to each mapped domain. Based on the identified problems and strengths, the Problem Rating Scale for Outcomes is assigned a Knowledge, Behavior, and Status score, which leads to the Omaha System Intervention scheme. Subsequently, an intervention string combined with the three hierarchical levels of the Intervention scheme is created to capture evidence of effectiveness of what works and trigger solutions which promote positive attributes and health outcomes of individuals, populations, communities, and organizations.

Results and Discussion

Following the strengths-oriented framework and the Omaha System pathway schemata, an open source Strengths-Oriented Diabetes Care Plan was created by mapping to the 2019 standards of diabetes care, literature, and expert opinions. The Omaha System domains and problem concepts captured in the diabetes care plan are illustrated in Figure 3. This diabetes care plan aims to capture optimal diabetes management, which can be achieved by collaborative and systematic engagement of a coordinated team of health care professionals to provide patient-
centered high-quality care based on evidence-based guidelines. Research and evidence also support that optimal diabetes management should also assess and address psychosocial issues and decrease or eliminate financial barriers for diabetes care and management. On the community and organizational levels, optimal diabetes management is supported by identifying, developing, and engaging community resources and public policies that support healthy lifestyles. Social determinants of health are defined as economic, environmental, political, and social conditions in which people live. They have contributed significantly to health inequality worldwide and exert a big impact on health inequities related to diabetes and its complications. This diabetes care plan incorporates all these comprehensive factors and resources on all the various levels.

Figure 3. The Omaha System domains and problem concepts by number.

Note: Env = Environmental; Phy = Physiological; Psy = Psychosocial; HRB = Health-related Behaviors

Incorporating an evaluation scale allows multidisciplinary teams to track disease management and health promotion strategies and measure health-related behaviors and outcomes at both individual and collective levels so that the care intervention process and tools can be modified for better care qualities and results. The care intervention process involves 4 actionable categories in different proportions (Figure 4) of care activities that focus on a total of 31 unique targets (Figure 5).

Figure 4. The Omaha System categories by number.
Figure 5. The Omaha System targets by number.

An intervention string was constructed by combining a problem term from the Problem Classification Scheme, a category, a target, and a customizable care description from the Intervention Scheme. This intervention string delineates an intervention activity or strategy targeted to optimal diabetes care management and health promotion. Examples of the intervention strings from this care plan are displayed in Table 1. These intervention activities subsequently lead back to an iterative process to reevaluate previous identified problems and strengths and associated outcomes that warrant the sustainment or modification of existing interventions or development of new interventions to address problems and strengthen positive attributes and assets.

Table 1. Examples of Omaha System diabetes care plan in English and Chinese versions.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Category</th>
<th>Target</th>
<th>Care Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health care supervision</td>
<td>CM</td>
<td>continuity of care</td>
<td>patient-centered collaborative care and tailored treatment for social context</td>
</tr>
<tr>
<td>Health care supervision</td>
<td>S</td>
<td>medical/dental care</td>
<td>ongoing assessment of pain, circulation, neuromusculoskeletal function, vision, and mental health by the health</td>
</tr>
<tr>
<td>Health care supervision</td>
<td>TGC</td>
<td>medical/dental care</td>
<td>multifactorial risk-reduction strategies beyond glycemic control</td>
</tr>
<tr>
<td>Medication regimen</td>
<td>CM</td>
<td>medication/ordering</td>
<td>resources to obtain needed medications</td>
</tr>
<tr>
<td>Medication regimen</td>
<td>TP</td>
<td>specimen collection</td>
<td>medication-specific diagnostic test</td>
</tr>
<tr>
<td>Nutrition</td>
<td>S</td>
<td>dietary management</td>
<td>follows recommended nutritional plan</td>
</tr>
<tr>
<td>Nutrition</td>
<td>TCG</td>
<td>dietary management</td>
<td>hypoglycemia treatment: eat 15 grams of carb, wait 15 minutes, retest, if still low, repeat treatment</td>
</tr>
<tr>
<td>Physical activity</td>
<td>TGC</td>
<td>wellness</td>
<td>integrate physical activity into daily routine</td>
</tr>
<tr>
<td>Skin</td>
<td>CM</td>
<td>personal hygiene</td>
<td>refer to outpatient foot clinic for ongoing nail and foot care as needed</td>
</tr>
</tbody>
</table>
### Standardized Translation Model

To advance health promotion and disease management in a global setting, a process should be considered to engage standardization of the translation process to disseminate best practices in a global environment. Using a standardized terminology and language such as the Omaha System, health data and information can become easily interoperable, and readily comparable and sharable. Many evidence- and practice-based care plans for management of various chronic populations in community developed by using the Omaha System have been integrated into health information systems that can be disseminated into the global communities. When initiating the translation process of electronic tools for diabetes management within the global context of health management exchange, a standardized translation model (Figure 6) for vocabulary mapping should be taken into consideration to harmonize different languages and achieve a shared purpose of information representation. The translation process of the strengths-oriented diabetes care plan into Chinese utilizes this proposed conceptual translation model.

In this model, a special intermediate tool such as the use of a middleware including standardized terminologies of SNOMED CT, the Omaha System, and others should be used to convert the vocabularies of these systems and align them in one congruent context based on multiple constructs of a vocabulary mapping that harmonizes the two languages with standardization. Such knowledge modeling that provides common data elements and terminology can enable the representation of data, semantics, and knowledge in healthcare information over time and location independent from technology. Specific constructs should be put into the context of lexical, semantic, conceptual, process, and cultural considerations. This way, health management tools can become a standardized data and information exchange resource and discourse for health management and outcome measurement in the community within a global health context.
Figure 6. Internationalization Architecture of Omaha System.

Conclusion

This project demonstrates a novel strengths-oriented interdisciplinary approach to disseminate best practices of diabetes management in global health communities using a standardized multidisciplinary terminology. It offers the potential to bring person-centered collaborative care to multi-levels of engagement that generate actionable and measurable solutions based on both problems and assets of strengths and resources. Such collaboration opens a continued dialogue in the discourse for leveraging the use of global health informatics principles and practice to reduce the burden of diseases around the world.

Competing Interests Statement

The authors have no competing interests to declare.

Contributorship Statement

This paper was composed by the first author and reviewed, edited, and approved by co-authors.

References

15. Martin KS. The Omaha System: A Key to Practice, Documentation, and Information Management (Reprinted 2nd ed.). Omaha, NE: Health Connections Press; 2005.


genoDraw: A Web Tool for Developing Pedigree Diagrams Using the Standardized Human Pedigree Nomenclature Integrated with Biomedical Vocabularies

Luciano Garcia-Giordano1, Sergio Paraiso-Medina, PhD1, Raul Alonso-Calvo, PhD1, Francisco Javier Fernández-Martínez, PhD2, Victor Maojo, MD, PhD1,
1Biomedical Informatics Group, DIA & DLSIIS, ETSI Informáticos, Universidad Politécnica de Madrid, Boadilla del Monte, Spain; 2Genetics and Inheritance Research Group, Instituto de Investigación Hospital Universitario 12 de Octubre (i+12), Madrid, Spain

Abstract

The integration of genetic information in current clinical routine has raised a need for tools to exploit family genetic knowledge. On the clinical side, an application for managing and visualizing pedigree diagrams could provide genetics specialists with an integrated environment with potential positive impact on their current practice. This article presents a web tool (genoDraw) that provides clinical practitioners with the ability to create, maintain and visualize patients' and their families' information in the form of pedigree diagrams. genoDraw implements a graph-based three-step process for generating diagrams according to a de facto standard in the area and clinical terminologies. It also complies with five characteristics identified as indispensable for the next-generation of pedigree drawing software: comprehensiveness, data-drivenness, automation, interactivity and compatibility with biomedical vocabularies. The platform was implemented and tested, confirming its potential interest to clinical routine.

Introduction

Informatics tools able to exploit genetic information in familial inheritance are increasingly necessary in clinical practice. Making sense of the genetic relations among individuals is an important task even in day-to-day medical activities in the genetics area. However, the availability of applications that allow for registering and visualizing family genetic information is not yet up to par with the necessary requirements for the task. In our long-term collaboration with the 12 de Octubre Hospital, Madrid, Spain, genetics specialists showed an interest in having a tool to facilitate the creation, management and visualization of pedigree diagrams, a visual system to represent families, in the day-to-day clinical activities of these specialists. However, the requirements presented to us for such a tool revealed that some characteristics essential for pedigree drawing systems to be of use in clinical practice are not found in the currently available set of platforms for this purpose. The first characteristic should be (a) that the system complies with the Standardized Human Pedigree Nomenclature – a de facto standard - in its updated version and is able to represent even the less-common scenarios that can be of interest for being documented; (b) that the tool is capable of automating the process of drawing the pedigree; (c) that the system is capable of generating the diagram from structured data. Having the characteristics (b) and (c), the diagrams do not need to be stored as visual diagrams (as images, for instance), but as structured data of each individual represented, as well as relations among them. Thus, data retrieved from medical information systems can be represented as pedigree diagrams with minimal intervention of the user. One of the necessities for this tool was that it was easy to use in a clinical scenario. For this purpose, we concluded that another key characteristic should be (d) that the system would present good usability characteristics, enabling the user to interact easily and rapidly with the interface. Thus, specialists could use the tool to create and manage pedigrees during their medical encounters with patients. Additionally, since the tool must be adapted to work in a clinical environment, some integration aspects should be considered. In our analysis, one integration capability that was especially important was that (e) the individuals represented in our tools should have their traits and diseases annotated as terms extracted from widely-adopted biomedical vocabularies (i.e. Human Phenotype Ontology (HPO), SNOMED-CT and the Online Mendelian Inheritance in Man (OMIM)).

In this paper, we refer to the five described characteristics as (a) comprehensiveness, (b) automation, (c) data-drivenness, (d) interactivity, and (e) compatibility. They are particularly important in the context of precision medicine, in which very diverse situations must be represented, and data stored in clinical information systems must be retrieved and shown to the specialist. Then, specialists will be able to take more information into consideration when making,
for example, a diagnosis, or provide a better, more personalized treatment. Additionally, the presented characteristics offer the possibility to better analyze the way in which families are structured and how genetic diseases are inherited, opening possibilities in statistical and medical research.

An analysis of the pedigree diagram drawing tools and platforms currently available revealed that some of the five characteristics can be found in some of these platforms. Examples of such platforms are Madeline 2.0, My Family Health Portrait, CRA Health, Progeny, and GenoPro. However, none of the tools analyzed combine the five characteristics simultaneously. Furthermore, none of the tools implement the updated version of the nomenclature mentioned above, and, to our knowledge, none of the systems support the annotation of diseases as terms from biomedical vocabularies. In order to offer a solution that aims to comply with the five characteristics identified, we developed genoDraw (www.genodraw.com). In this work, we present the foundations on which this system is built.

Methods

We identified the five characteristics that we believe are necessary for a better acceptance of this kind of system in clinical practice. Hence, the foundational idea around genoDraw is to be a platform for the creation, management and visualization of pedigree diagrams. In this section, we present the basis of our approach for the representation engine and the strategies that we applied so that each of the characteristics (a-e) is addressed by genoDraw. In our specific case, a web-based platform was ideal. At the end of this section, we present the architectural aspects of our system and other implementation details.

First, to represent the pedigree diagrams, we follow the Standardized Human Pedigree Nomenclature. This nomenclature is a recommendation of the National Society of Genetic Counselors and it is a current de facto standard in the discipline. The definitions and rules established by the updated version of this nomenclature enable us to define an algorithmic process for the diagram creation in a broad range of reproductive scenarios. Examples include planned adoptions, ovum donations, surrogate gestations, as well as the expected usual situations, such as single and multiple gestations.

Second, to correctly store the necessary entities and links and draw them on a canvas, we represent the diagram internally as a graph (structured data). Entities (nodes) of this internal graph are individuals, gestations and relationships. In each of them, data are stored. For instance, a gestation can be monozygotic or not when more than one child is listed, and an individual might be affected by a certain trait. A second graph, the one to be drawn as a pedigree diagram, is generated using data describing these entities and their relations (biological parenthood, partner of a relationship, etc.). This is accomplished using data describing these entities and their relations (biological parenthood, partner of a relationship, etc.). This is accomplished using the three-step process described below.

The three-step process involved in drawing pedigree diagrams is responsible for transforming the internal graph, which stores what is essential to the representation as structured data, into the representation graph. The representation graph, in turn, follows the nomenclature and does not necessarily contain the same nodes and links as the internal one. In fact, in many scenarios, the representation graph will be composed of more nodes and fewer edges than the internal one. For example, the internal graph stores, for each gestation, each of the parents. However, when a relationship between the parents is identified, only one link between the relationship and the gestation is drawn, and not those between the gestation and each parent. As an illustration of this example, Figure 1 shows, on the left-hand side, the objects of the internal graph relative to three individuals (A, B and C), a relationship (R) and the gestation of C (G), as well as the connections among them. The resulting nodes and links of the matching pedigree diagram, drawn after following the three steps, are shown on the figure’s right-hand side.

The first step of the three-step process generates each of the nodes to be drawn. This step decides which entities are to be drawn or not and draws the squares, circles, diamonds and other artifacts that visually represent each entity. For individuals, characteristics such as gender, being deceased or not, and traits by which they are affected are characteristics that are reflected in the symbol drawn on the canvas. In Figure 1, for instance, the individual C is affected by sickle cell anemia (OMIM: #603903). This disease is annotated as a term from the Online Mendelian Inheritance in Man (OMIM) and, as described later, is included in a collection of traits of the corresponding individual as a reference to such term in the OMIM vocabulary. This characteristic is reflected in the representation graph as a gray mark in the node that corresponds to the individual C. In the case of gestations, only those that correspond to multiple gestations will have drawn at the node other than the gray circle, which we add for usability purposes as discussed later. When
Figure 1: A simple example of generation of a pedigree diagram from an internal graph. The internal graph is represented visually on the left side of the image, and the representation graph on the right side. As we can see, the links between the gestation and each parent are substituted by one link between the gestation and the relationship of the parents. This simplification is only done when possible. In accordance with the adopted nomenclature, an empty circle corresponds to a female individual and an empty square to a male. The horizontal line between $A$ and $B$ is their relationship, and the vertical line symbolizes that both are parents of $C$. The gray circles over the relationship and gestation nodes are added for usability purposes.

A node refers to a relationship, only cases of infertility and divorce are annotated visually. Otherwise, only the gray circle is drawn.

The second step generates the edges between represented nodes. Visually, they are the straight lines that connect related entities in the pedigree diagram. This step is based on rules derived from the nomenclature, and explores the internal graph, searching for specific patterns that correspond to drawing a connection between two nodes of the representation graph. Thus, this is the step that makes sense of the relations between nodes. A person is to be connected with his or her gestation if such gestation is decided to be drawn by the previous step. Similarly, a gestation is connected with each parent, or their relationship, if there is one. A relationship, in turn, is to be connected with both partners. By applying the previously mentioned rules, from a given situation we are able to reach a representation graph that is isomorphic to the pedigree derived directly from following the directives of the nomenclature we adopted, with some extensions that will be discussed later.

The third and last step for displaying a pedigree diagram as a graph is the definition of rules for the positioning of the nodes. In our system, the positions of the nodes are calculated following an optimization process conducted after the three steps. One of the inputs of this optimization process is information about which rules need to be vertically or horizontally aligned. Similarly to a solution proposed elsewhere\cite{11}, we define linear constraints for each of these alignments. The third step is the step which generates such linear constraints. This is done by exploring the internal graph in a similar manner as in the second step (by searching for specific patterns that activate rules for the relative positioning of the nodes). An activation triggers the creation of constraints that are followed by the optimization process. One example of alignment rule is that a node of type gestation is to be vertically aligned to the child if there is only one child. As it is depicted in the right side of Figure 1, the node corresponding to the gestation of the individual $C$ is vertically aligned to the node corresponding to the individual $C$.

After the third step, we obtain a representation of the pedigree diagram. The nodes are drawn on a canvas visible to the user, the connections (edges) between the nodes are also defined and drawn, and the constraints for alignments are defined. From this point on, the representation is handled by an engine specialized at arranging the nodes on the canvas according to an optimization algorithm and displaying this representation to the user. The optimization algorithm outputs positions for each of the nodes complying with the constraints defined. The optimization is, therefore, the minimization of a stress function with constraints. The function measures the difference between an ideal distance and the current distance between two nodes directly connected. This ideal distance is calculated for each edge of the graph and depends on the types of the nodes it connects. The minimization process is done iteratively and is repeated whenever a change to the representation graph is done. For example, when the user moves a node to another position, the minimization is triggered, changing iteratively the positions of the nodes until a new convergence is found. For
the specific purpose of this module, we implement a solution based on an already-existing library\textsuperscript{12} that provides an optimization algorithm that is useful for our necessities and enables the user to interact with the representation graph by moving the nodes.

As an illustration of the whole three-step process, we include Figure 2, in which the individual $B$ is affected by, for example, sickle cell anemia (OMIM code: #603903), has a relationship with $A$ and has had one with $C$. With individual $C$, $B$ had twins. We do not know if the gestation was monozygotic, hence the question mark (?) below the gestation node. Both daughters are affected by the disease. The individual $B$ now has a relationship with $A$, and they adopted $F$. The dashed line is drawn as so because, since the child is adopted, there is a nonbiological relationship between each parent and the adopted child.

![Figure 2](image)

Figure 2: Example of the three-step generation and arrangement processes. On the top-left are nine generated nodes that are the output of the first step of our process, which defines which nodes are to be drawn and draws them according to the nomenclature, with a few stylistic changes. Next, the second step defines the connections between nodes, which are the gray lines drawn. After that, the third step defines the constraints that the optimization engine uses to arrange the nodes on the canvas.

In terms of the characteristics we aimed for our system to possess, by having followed the updated Standardized Human Pedigree Nomenclature\textsuperscript{1}, we can ensure that our system can be, in principle, comprehensive enough to represent all the major scenarios that exist in our society. However, we detected some limitations that violated this ideal. One of the limitations is that, according to the symbology of the nomenclature, it is not explicit how a gestation with multiple children should be represented in the case of gestational surrogacy. Another limitation is that, if two individuals that have a relationship have family members in common, they might be included in different generations, thus, horizontally aligning their relationship edges is not feasible. In our system, we addressed these limitations by extending the nomenclature, making it more flexible when this would not render the pedigree diagram nomenclature ambiguous. Our minor changes provide more consistency and flexibility to the representation. In the case of the representation of multiple gestations, we chose to always show the gestation of a person when the parents of this person are also represented. That way, if it is the case of a multiple gestation, two people will be connected to the gestation. In the case of multiple gestation in the context of gestational surrogacy, for example, with our extension, there is no restriction to the child being only one. For the second limitation, in which the two partners of a relationship are members of different generations of the same family, our solution is the removal of the constraint of the partners being horizontally aligned only when there is a conflict in their generations. By implementing a system with our extended nomenclature, we ensure a wide comprehensiveness, by which the system is capable of representing all the situations that can be modeled following the directives of the nomenclature, which are most of the common situations that one can observe in the population.

Data-drivenness and automation are both addressed by defining our three-step process with the focus on generating
the diagram from structured data of individuals and their relations and by implementing a representation engine based on optimization\(^1\). Up to the end of the third step of our three-step process, the generation is fully automated. Having the nodes, the edges and the constraints (alignment rules) that define the pedigree, the representation engine arranges the nodes following an iterative optimization process. The full automation of the representation engine is not possible mainly for two reasons. The first is that structures reached by the optimization engine can be the result of convergences to local optima. That is, the ideal positioning of the nodes might not be reached. The second reason is that personal inclinations of the user are not considered during the positioning of the entities in the canvas. Both reasons might cause a need for small corrections and adjustments to the arrangement of the pedigree.

By allowing the user to move the nodes of the pedigree while keeping the constraints active, we address part of the fourth characteristic we defined as important for our system, which is **interactivity**. If the disposition of the nodes does not satisfy the user, he or she can change the positions of some nodes and a new iterative process is initiated, and convergence around the new arrangement is reached. One important characteristic of our representation interaction model is that, while a node is being moved, no rules of the disposition of pedigree diagrams are disregarded. That way, while allowing the user to make changes to the drawn diagram, our system ensures that the representation is correct. Another feature that also contributes to interactivity is the possibility to add and remove entities of the internal graph by interacting with nodes and context menus associated with each node of the pedigree. From the changes caused in the internal graph, another representation is generated and displayed. When possible, the positions of the existing nodes are kept, so that the user does not need to make any further changes to the parts of the diagram that were already arranged to their expectations. To edit information regarding a specific entity of the internal graph, we implemented a sidebar menu through which information about individuals, gestations and relationships can be inserted, changed or removed. Additionally, to provide better usability of the representation system, we represent the nodes that would correspond only to changes of directions between lines in a paper-based diagram with a semi-transparent gray circle. This feature enables the user to better understand with which entities of the canvas he or she can interact, either by dragging to define a new position or by clicking to view or edit the data associated with the entity that corresponds to the node clicked.

From a usability perspective, the user interacts with the canvas both directly and indirectly. Direct intervention is preferred for simple actions, such as the creation or elimination of entities. Adding a child to an individual, eliminating one or hiding it from the canvas are also direct interactions triggered using context menus directly on the corresponding nodes in the canvas. Indirect interaction is preferred for more complex actions, such as changing the name of an individual or adding another child to an existing gestation, which are made using a sidebar menu. Depicted in Figure 3 is the interface that allows managing a diagram. As we can see, the context menu in the center can trigger simple actions, while the sidebar menu on the right is used for more complex data changes.

The last characteristic we defined as being fundamental for a pedigree drawing software to include in the current context of data integration is **compatibility** with widely adopted biomedical vocabularies for the annotation of genetic traits and diseases. In this regard, we enable the user to assign traits to individuals in such a way that the traits are considered by the system as terms from biomedical vocabularies. Examples of vocabularies supported by genoDraw are HPO\(^2\), OMIM\(^4\) and SNOMED-CT\(^3\). Depending on license-related aspects, the selection of vocabularies can be changed. In this regard, genoDraw can operate on an unlimited number of vocabularies. This compatibility with biomedical vocabularies provides various advantages, such as common reference for each disease, and the possibility to develop further methods to make statistical analysis of the data inserted in internal graphs.

In terms of the **architecture** of the system, because genoDraw is intended as a platform for the creation, management and visualization of pedigree diagrams, a web-based system was ideal. In Figure 4, we describe the architecture of genoDraw. After creating a pedigree by interacting with our platform and inserting information of individuals and their relations, the user can save locally a file that includes the internal graph. Having this file, the user can load it on the platform to visualize the pedigree and make further changes to it.

As illustrated in Figure 4, genoDraw contains a backend that authenticates the user and provides the necessary files for genoDraw to work in the web browser of the user. The internal graph files are, in our implementation, JSON (JavaScript Object Notation) files managed by the user. An internal graph file contains the necessary information to generate a pedigree diagram (i.e. individuals, relationships, etc.). Such a file can be loaded into the platform but
Figure 3: Interface of genoDraw. The user is editing the example used in Figure 2 to add a partner to the individual F. As we can see in the sidebar menu, the relation between F and A and F and B are of type “nonbiological parenthood”. Other information available for the individual F (i.e. gender, name, fertility, etc.) is shown in the different sections of the sidebar menu, which can be extended and collapsed.

Figure 4: The architecture of genoDraw. The user interacts with a web client application connected with a backend that only handles the distribution of the website itself and the authentication of users. The internal graphs (structured data) are files which are stored locally in the user’s device. One detail of our implementation is that we make use of a database system to handle authentication information. A web server (backend) handles this authentication process and serves the files corresponding to genoDraw to the client. We implemented genoDraw in JavaScript, making use of D3.js and WebCola as libraries for handling the visualization. We selected WebCola as the graph-handling library due to
its capability to keep the alignment constraints active while the user manipulates the graph, as well as its capability to handle the structuring process of the diagram (optimization process) based only on alignment rules (constraints). Furthermore, it allows for an interactive use in a wide range of devices, of which traditional computer devices, such as desktops and laptops, as well as tablets, are the most used by our target users.

Results

We designed genoDraw to be adopted in clinical practice as a platform of intense, day-to-day use. Genetics specialists are expected to use our tool in medical encounters.

To evaluate the usability of our system, we have carried out a usability test. The objective of the test was to analyze the time required for a subject to be acquainted with the platform and to observe usability variables, such as if the user was capable of finding the way to insert or alter specific data in the diagram, and how long it took. Additionally, we obtained feedback from the users with some satisfaction questions. The subjects were 26 graduate students in the area of biomedical engineering who had been previously informed about the nomenclature but had no knowledge of our tool. We first let them explore the platform for a few minutes. Then, we presented the students two scenarios of different complexities in plain text and asked them to represent the scenarios as pedigree diagrams using genoDraw. The scenarios contained multiple gestations, ovum donations, surrogate gestations and adoptions, so as to observe what kinds of usability problems could exist in the insertion of the more complex and uncommon scenarios. Furthermore, we asked the students to insert traits and diseases of each individual as specific terms from standard vocabularies. The results of the usability test revealed that most users were familiarized with genoDraw in less than 30 minutes. We also observed that the time required for the users to represent both scenarios ranged from 20 minutes to 50 minutes. The average completion time was 20 minutes for the first scenario and 10 minutes for the second. In the first scenario, which contained a family with a case of multiple gestation, some of the users had difficulties finding the correct way of inserting such type of gestation. We interpret this as a usability issue, since the only path to add a multiple gestation at that moment was to include an additional child to an existing single gestation. In the second scenario, an ovum donation was to be represented, which meant that the users should insert the donor as biological mother in the gestation of the child. While most of the users found the path to adding such information in the pedigree without much effort, many agreed that it was not an intuitive solution. Having detected these and other usability issues, we addressed them.

To assess the correctness of the pedigree diagrams generated using our system, an evaluation of the tool was required to detect and address any incompatibilities or deviations from the standard nomenclature adopted for our system. In our case, we conducted this assessment by progressing through some case studies step-by-step. These practical scenarios were designed by experts at the Genetics Unit of the 12 de Octubre Hospital to serve as examples for our evaluation in plain text. After methodically introducing each of the examples in genoDraw, we then compared our results with the recommended results also given by the experts. During this evaluation phase, we detected some issues regarding the correctness of the tool and solved them.

An example of one of the validation cases is the following: A couple is formed by a man \(A\) and a woman \(B\). The woman is affected by retinoschisis in its X-linked recessive variant (OMIM code #312700) and has had, with the same man \(A\) with whom she forms a couple, one son \(C\) also affected by the disease, and is pregnant with a child \(D\). She has a brother \(E\) who is also affected and a sister \(F\) who is not. Her father \(G\) is also affected by the same disease and is married to his cousin \(H\), who is the mother of \(B, E\) and \(F\). Our result for this example is shown in Figure 5. Having a representation generated from a case such as the example described, a specialist is able to visualize the family, assessing the risk of the child in pregnancy being affected by the disease, as well as identifying other family members who might be, for example, carriers of the disease.

Discussion

The purpose behind genoDraw is to provide genetics specialists with a tool to address their needs. In this work, we identified these needs as five characteristics the tool must possess, and designed a tool that provides the resources necessary to create, manage and visualize pedigrees via its ability to provide solutions for each of the five characteristics. genoDraw is a tool intended to be used in clinical practice by medical specialists in the area of genetics. For this
reason, we carried out a thorough evaluation of our tool, focusing not only on its correctness, but also on its ease of use. With genoDraw’s current capabilities, a specialist can insert information during a medical encounter and observe specific characteristics of the family of the patient. This is undoubtedly an advantage of any pedigree representation, even when sketched on paper. However, since the pedigree diagrams in genoDraw are generated automatically from the data inserted, the user does not need to plan which entities should be drawn. That way, the composition of a pedigree diagram is as easy as a step-by-step insertion of information. Additionally, since the arrangement of the diagram is done automatically, no planning of where each entity should be drawn is needed. Thus, according to our observations, genoDraw is capable of enabling an improvement in clinical practice.

In terms of the limitations of our tool, some should be noted. The first is that, despite the fact that the representation engine is flexible, it is constrained by the bidimensionality of the canvas on which the diagram is drawn. Thus, issues such as the planarity of the graph being represented or the alignment impossibilities, described elsewhere, are unavoidable. Another limitation is the algorithmic procedure used to arrange the nodes of the graph in the canvas. Although the generated graph is always correct according to the nomenclature with our additional described changes, the structure is defined by a process that does not consider factors such as the personal appreciation of the user, who might want a specific arrangement of the nodes. In this case, we present a solution by enabling the user to make changes to such positions so that the optimization process finds a new convergence around the new structure. Further enhancements improving the tool’s ability to adapt to the user’s preferences should enable the generation of better-arranged pedigree diagrams.

Additionally, the presented platform is currently capable of storing the individuals’ diseases and traits as terms from widely used biomedical vocabularies. Although this is a novel feature, it does not yet enable full integration of the system with other, currently used, systems. We intend now to expand the compatibility of our system towards the integration with the represented individuals’ electronic medical records. This could make of our system a tool that retrieves, updates and uses other relevant information stored in these records to enhance and facilitate the risk assessment and diagnosis of genetic diseases. A future development in this direction will be the expansion of the system with the capability of generating clinical messages that can be stored in clinical environments, thus integrating the tool with other currently used clinical data storage systems.

When compared to other tools that serve the same purpose, we believe genoDraw is uniquely suited. In terms of comprehensiveness, our tool is the only system reported in the literature that complies with all the directives established in the updated version of the Standardized Human Pedigree Nomenclature. In this regard, the flexibility of our model is key to this characteristic. Another solution, for example, considers that every individual is child of a relationship, not being flexible enough to register that one’s biological parents might not have any relationship, such as in the case of sperm donations. As far as interactivity is concerned, our solution for arranging and moving the entities of the pedigree on the canvas enables the user to interact with the diagram without rendering it incorrect. According to our observations, this is superior to other solutions proposed. In GenoPro and in the online version of Progeny, for
example, the structure of the pedigree can be voided by moving nodes. In terms of compatibility with biomedical vocabularies, to our knowledge, genoDraw is the only tool that enables the annotation of diseases as terms from biomedical vocabularies. Lastly, the automation and data-drivenness provided by genoDraw are enough to generate a plausible pedigree diagram for all situations, while other tools that generate pedigree diagrams from data tend to present various limitations in this regard.

Conclusion

In this work, we present genoDraw, a platform for representing, creating and managing pedigree diagrams, which is built around five characteristics that we identified as necessary for the next generation of such tools. The central module of this platform is a representation engine, which is capable of deriving a pedigree diagram from structured data. In order to integrate the system with other medical resources, our implementation is also compatible with annotations of diseases and other traits as terms from widely adopted biomedical vocabularies, in our case SNOMED-CT, OMIM and HPO. Depending on licensing aspects, the selection of available vocabularies can be changed, and many vocabularies are compatible with genoDraw. Our system supports the visualization and creation of pedigree diagrams from and to structured information that can come to be stored in medical information management systems, and the diagrams are generated following the Standardized Human Pedigree Nomenclature in its updated version. Thus, genoDraw is a tool that can improve the current practice of sketching pedigrees on paper by providing genetics specialists with useful computing capabilities that include the ability to conveniently input information during a medical encounter, the fact that this input data is kept as structured data that can be used to update the information of patients, as well as the ability to share internal graph files with other health professionals in order to provide a better, more precise treatment.

In the near future, we intend to integrate genoDraw into the day-to-day workflow of specialists in genetics and other areas for the assessment of the genetic component of various illnesses in patients. In terms of its use in research, we envision genoDraw as being a tool with the potential to support data collection of individuals and their families.

Acknowledgment

This work is supported by “Proyecto colaborativo de integración de datos genómicos (CICLOGEN)” PI17/01561 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2017-2020 and the European Regional Development Funds (FEDER).

References

7. CRA Health suite. (n.d.). CRA Health LLC.


Deep Learning from Incomplete Data:
Detecting Imminent Risk of Hospital-acquired Pneumonia in ICU Patients
Travis R. Goodwin, PhD, and Dina Demner-Fushman, MD, PhD
Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
Bethesda, MD, USA

Abstract
Hospital acquired pneumonia (HAP) is the second most common nosocomial infection in the ICU and costs an estimated $3.1 billion annually. The ability to predict HAP could improve patient outcomes and reduce costs. Traditional pneumonia risk prediction models rely on a small number of hand-chosen signs and symptoms and have been shown to poorly discriminate between low and high risk individuals. Consequently, we wanted to investigate whether modern data-driven techniques applied to respective pneumonia cohorts could provide more robust and discriminative prognostication of pneumonia risk. In this paper we present a deep learning system for predicting imminent pneumonia risk one or more days into the future using clinical observations documented in ICU notes for an at-risk population (n = 1,467). We show how the system can be trained without direct supervision or feature engineering from sparse, noisy, and limited data to predict future pneumonia risk with 96% Sensitivity, 72% AUC, and 80% F1-measure, outperforming SVM approaches using the same features by 20% Accuracy (relative; 12% absolute).

Introduction
The Centers for Disease Control (CDC) estimates that one in every twenty-five acute care hospitalizations results in a health care-associated infection (HAI). Recent analyses indicates that at least 50% of HAIs are preventable. Not only are HAI used to measure quality of care by the Centers for Medicare and Medicaid Services (CMS), failure to prevent HAIs can result in financial penalties such as those described in the 2010 Patient Protection and Affordable Care Act. Moreover, HAIs are estimated to cost over $9.8 billion USD annually, with hospital acquired pneumonia (HAP) costing an estimated $3.1 billion alone. Hospital acquired pneumonia is the second most common nosocomial infection in the intensive care unit (ICU) and the most common in mechanically ventilated patients. Hospital and community acquired pneumonia require different treatment plans (i.e., antibiotics) than other lower respiratory tract infections. Clearly, the ability to predict pneumonia risk can improve patient outcomes by enabling early interventions, monitoring and improved clinical decision support.

A number of pneumonia risk prediction models have been developed which rely on hand-curated signs and symptoms. In an external evaluation of these models, Schierenberg et. al. (2016) found that existing pneumonia models typically distinguish between risk and no risk, but fail to discriminate between high- and low-risk patients. Moreover, the majority of models target community acquired pneumonia, with the few HAP prediction models being focused on predicting the risk of HAP following a specific procedure. Consequently, we were interested in evaluating whether data-driven techniques enabled by recent advances in deep learning would enable us to automatically predict when at-risk individuals progress to high- (i.e., “imminent”) risk by considering a richer set of clinical observations than those used in prior pneumonia risk prediction models. Specifically, we were interested in harnessing the clinical narratives documented in ICU notes. Clinical narratives provide a rich but underutilized source of information for clinical decision support, potentially documenting and/or summarizing the main observations about the patient, relevant procedures and important positive and/or negative laboratory results. Unfortunately, processing clinical narratives requires overcoming several barriers including the prevalence of missing, inconsistent, or underspecified information. Moreover, while typical risk predictors such as vital signs are continually recorded through a patient’s stay, clinical notes are produced at irregular intervals and often in bursts (that is, there are often multiple days during a patients ICU stay in which no clinical notes are produced). Consequently, inferring pneumonia-risk from clinical notes requires accounting for (1) incomplete or sparse information, (2) gaps in the patient’s time line in which no notes were generated, (3) limited availability of training data, and (4) lack of direct (ground-truth) pneumonia risk labels.

In this paper, we present the Pneumonia Risk predictiOn NeTwOrk (PRONTO) which harnesses modern deep learning techniques to infer and predict when and if an at-risk ICU patient will progress to imminent pneumonia risk within 467
a given time window based on the content of his or her longitudinal ICU notes. Specifically, we identified a cohort of at-risk ICU patients \( n = 1,467 \) who developed pneumonia during their stay and used an emergent deep learning technique known as Recurrent Additive Networks\(^{11}\) (RANs) to jointly predict the progression of pneumonia risk, relevant clinical observations, and temporal interactions based on longitudinal analysis of ICU notes generated for a retrospective pneumonia cohort. We show that not only can PRONTO be successfully trained from a limited, sparse (i.e., incomplete), and noisy data set, but that it substantially outperforms SVM-based alternatives by up to 20\% (relative; 12\% absolute) increased accuracy.

**Background and Significance**

Classic models of disease-risk prediction have relied on a small set of specific risk factors. For example, the Heckerling Clinical Decision Rule for the Diagnosis of Pneumonia evaluates the risk for Pneumonia based on five binary risk factors\(^5\) indicating whether the patient’s (1) temperature is above 37.8°C or (2) heart rate is above 100 bpm as well as whether or not the patient has (3) crackles/rales, (4) decreased breath sounds, or (5) asthma. By comparison, PRONTO considers every risk factor or clinical observation recorded in each clinical note. Consequently, it can be seen as a generalization of these models, though we note that some routine information like temperature may not always be documented in the narrative.

Previous work for disease-risk prediction using natural language processing has largely focused on a small set of features and on classifying risk at the time that each clinical note was written. For example, Bejan et al. \((2013)^{12}\) represented patient’s clinical history as a sequence of days (with no gaps) where each day was associated with a clinical note. They designed a support vector machine (SVM) for classifying whether a patient is positive or negative for pneumonia at each day based on the clinical notes produced up-to and on that day. By contrast, PRONTO allows for gaps in a patient’s record and can predict whether the patient will be positive or negative for pneumonia at given number of days in the patient’s future for which no clinical notes (yet) exist.

Risk prediction at arbitrary times was considered in Goodwin and Harabagiu \((2015)^{13}\) in which the authors constructed a multi-layer Hidden Markov Model to predict the presence or absence of seven risk factors associated with heart failure based on the previous clinical note. Unfortunately, that methodology requires pre-computing co-occurrence information for all observations, which does not scale well to large numbers of observations. PRONTO, by contrast, is able to consider a large number of unique observations (28,782 in our experiments), and can predict the risk that a patient will develop pneumonia in the future.

**Materials**

Our experiments relied on MIMIC-III\(^{14,15}\), a publicly-available critical care database developed by the Massachusetts Institute of Technology (MIT) Lab for Computation Physiology to support research in intelligent patient monitoring. MIMIC-III contains de-identified health data associated with over 40,000 patients. Although MIMIC provides a wealth of structured information including demographics, charts, laboratory tests, medications, and diagnoses, we exclusively relied on the unstructured textual data (i.e., clinical notes). We obtained an initial retrospective cohort of 281,076 patients in MIMIC-III who were coded with a discharge ICD-9 diagnoses indicating community or hospital acquired pneumonia (ICD-9 486). Within this cohort, we considered only ICU stays in which the patient (1) was not admitted with pneumonia (i.e., pneumonia was not community acquired) and (2) was not diagnosed with pneumonia during the first day of their stay (using the approach for detecting pneumonia offset described in subsection C of the Methods section). This resulted in 1,494 unique ICU stay\(^{16}\) for 1,467 patients, with each stay associated with multiple clinical notes generated over multiple days. We created sub-cohorts for training, development, and testing using an 8:1:1 random split (at the patient level).

**Methods**

To account for the sparse or incomplete nature of ICU notes – both in terms of the observations documented within the note and in terms of the gaps in a patient’s stay without any notes – we considered an abstract representation of the patient’s ICU stay. Specifically we discretized the the patient’s ICU stay into discrete, non-overlapping, discontinuous 24-hour windows in which clinical notes were produced. We refer to information documented in each 24-hour window as a **clinical snapshot**, and to the discontinuous sequence of clinical snapshots during a patient’s ICU stay as his or her **clinical chronology** (details are provided in subsection B).

\(^{1}\)Note: we merged re-admissions within a 30 day window into a single ICU stay.
To overcome the noisy and incomplete information associated with clinical narratives we needed to design a predictive model which could: (1) discriminate between relevant and irrelevant clinical observations extracted from clinical notes, (2) learn latent temporal interactions between relevant clinical observations documented on different dates, (3) infer long-distance causal and inhibitive relations between clinical observations, their interactions, and the course of pneumonia. To this end, we present a deep-learning architecture named PRONTO (Pneumonia Risk predictiOn NeTwOrk).

**Figure 1** illustrates our three-step approach for pneumonia risk prediction:

**Step 1:** we automatically extract *clinical chronologies* for each patient in our training cohort;

**Step 2:** we train PRONTO to infer the risk of developing pneumonia based on latent temporal, causal, and inhibitive relations encoded in the clinical chronologies extracted for the retrospective pneumonia cohort described above; and

**Step 3:** we apply PRONTO to predict the risk of pneumonia within an arbitrary temporal window for a given patient based on his or her ICU notes.

Below, we (A) define and describe how we automatically extract clinical chronologies from longitudinal ICU notes, (B) present the architecture of PRONTO, (C) describe how the model can be trained from a retrospective patient cohort without direct supervision, and (D) explain how PRONTO can be used to predict pneumonia risk for new patients.

### A. Extracting Clinical Chronologies

Extracting the clinical chronology of a patient’s ICU stay from clinical notes requires overcoming several barriers. First, both the type of note (e.g., nursing note, radiology, admission report) as well as the frequency of notes varies from day-to-day, admission-to-admission, and patient-to-patient. Second, each note documents a different and incomplete (i.e., sparse) set of observations about the patient’s clinical picture at the time the note was written (with different types of notes emphasizing different parts of the patient’s clinical picture). Third, there are gaps in the patient’s ICU stay in which no notes are generated, and days on which multiple notes are generated.

Consequently we represent a patient’s ICU stay as a sparse, discontiguous sequence of *clinical snapshots* which are defined as the sets of observations about the patient’s clinical picture documented across any ICU notes produced on the same date. Thus, extracting the clinical chronology for a patient’s ICU stay reduces to (1) identifying the clinical observations reported in his or her ICU notes, and (2) organizing the observations into a sequence of clinical snapshots.

**Identifying Clinical Observations.** Clinical observations are often documented as multi-word nouns (e.g., “ventilator associated pneumonia”). Consequently, to identify clinical observations we first pre-processed each clinical note using the OpenNLP sentence splitter, tokenizer, lemmatizer, part-of-speech tagger, and dependency parser. After pre-processing, we identified clinical observations using MetaMap Lite and discarded observations whose semantic types did not correspond to problems, interventions, medications, anatomy, or findings. To detect other attributes, we developed a minor extension to MetaMap Lite based on FastContext, a high-performance re-implementation of ConText. FastContext associated each observation with the following semantic attributes:

- *negation* indicating whether the observation was affirmed or negated in the narrative;
- *certainty* indicating whether the author was certain or uncertain about the observation;
- *temporality* indicating whether the observation occurred in the present, the past, or is hypothetical; and
- *experiencer* indicating whether the observation was associated with the patient or someone else (e.g., family).

We considered only clinical observations which were affirmed, certain, present, and associated with the patient.

[https://opennlp.apache.org/](https://opennlp.apache.org/)
Figure 2: Clinical chronology extracted for a single patient’s ICU stay with 7 clinical snapshots.

**Mathematical Representation.** We represent the clinical chronology for a patient with \( L \) clinical snapshots as a pair \( C = (\delta, O) \) where \( \delta \in \mathbb{Z}^L \) is an elapsed time vector such that \( \delta_i \) indicates the elapsed time between the \( i \)th and \((i - 1)\)th clinical snapshots (where \( \delta_0 := 0 \)), \( O \in [1, V]^{L \times N} \) represents the clinical observation matrix such that each column \( O_i \) indicates the sequence of clinical observations documented in the \( i \)th clinical snapshot, \( V \) is the number of unique clinical observations documented in any clinical note associated with the training cohort, \( N \) is the maximum number of observations considered for any clinical snapshot, and \( L \) is the maximum number of clinical snapshots used for any clinical chronology. In our experiments, \( V = 28,782 \), \( N = 256 \), and \( L = 7 \).

**Figure 3:** Architecture of the Pneumonia Risk predictiOn NeTwOrk (PRONTO), where \( C \) indicates the clinical chronology extracted from a patient’s ICU stay, \( \delta \) indicates the prediction window, and \( \bowtie \) indicates vector concatenation.

**B. Architecture of the PRONTO Model**

Given a clinical chronology \( C \) and a prediction window \( \delta \), PRONTO predicts the risk the patient will develop pneumonia during that window using three phases illustrated in Figure 3:

- **Phase I: Clinical Snapshot Encoding** processes the sequence of clinical observations \( O_i \) in each clinical snapshot \( i \in [1, L] \) to produce an embedded representation \( d_i \) of the snapshot;
- **Phase II: Clinical Picture Inference** processes the encoded clinical snapshots along with their elapsed times to identify latent temporal, causal, and inhibitory interactions and infer the clinical picture of the patient at the end of the given prediction window \( \delta \); and
- **Phase III: Pneumonia Risk Prediction** relies on the inferred clinical picture to predict the probability that the patient will develop pneumonia within \( \delta \) days.

These three phases are detailed in the remainder of this section.

*To keep observations and the elapsed time within comparable domains (i.e., between 0 and 1), we compute the log of the number of elapsed days and project it between 0 and 1 using the hyperbolic tangent function.*

470
Phase I: Clinical Snapshot Encoding. The goal of the first phase of PRONTO is to learn an optimal encoding of individual clinical snapshots. Formally, we represent each clinical observation as a $V$-length one-hot vector. We designed PRONTO to use a Deep Averaging Network (DAN) for this purpose. However, due to the paucity in the literature for encoding sequences of sparse observations, we implemented and evaluated a total of five neural network architectures for encoding clinical snapshots, as shown in Figure 4. We first present well-known architectures for embedding documents and input sequences: Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).

RNN Encoder. The RNN architecture illustrated in Figure 4a operates by (1) learning an embedded representation of each clinical observation $O_{i1}; O_{i2}; \ldots; O_{i3}$ and (2) sequentially applying a forward RNN to learn how to compose, combine, or aggregate the embedded clinical observations to produce a single encoding for the entire clinical snapshot by maintaining and updating an internal memory vector. Note: as indicated in Figure 4a, we discard the individual outputs of the RNN and use the final output of the RNN as the clinical snapshot, i.e., $d_i := m_N$. In our experiments, we used Long Short-Term Memory (LSTM) units to build our RNN.

CNN Encoder. Like the RNN encoder, the CNN shown in Figure 4b initially learns an embedded representation of each clinical observation. The CNN then combines the embedded representations of adjacent clinical observations using three parallel convolutional filters, each followed by a maximum pooling layer. The results of the maximum pooling layers are combined (i.e., concatenated) together and projected into a snapshot embedding using a dense (i.e., fully-connected) Rectified Linear Unit (ReLU). As in Kim (2014), we used three filters operating on sequences of 3, 4, and 5 clinical observations with 1,000 kernels.

DAN Encoder. Less well-known than the RNN and CNN, the Deep Averaging Network (DAN) illustrated in Figure 4c also initially learns an embedding for each clinical observation. However, unlike the RNN and CNN, the DAN encoder does not consider the order of clinical observations within a clinical snapshot. The DAN encoder relies on $k$ dense layers to refine the clinical observation embeddings followed by an element-wise average, and $j$ dense layers to refine
the average embedding and produce the encoding of the clinical snapshot. As in Goodwin et al. (2017) we used $k := 2$ and $j := 2$ and ReLU activations on all dense layers.

**Dense Encoder.** Like the DAN, the dense projection encoder illustrated in Figure 4d ignores the order of clinical observations in each clinical snapshot. Unlike the DAN, however, the Dense encoder does not learn embeddings for individual clinical observations. Instead, the one-hot representations of each clinical observation are combined using bit-wise union (binary addition) to produce a single, sparse “bag-of-observations” vector. The bag-of-observations is directly embedded to produce the encoding of the clinical snapshot.

**Sparse Encoder.** Finally, the sparse encoder illustrated in Figure 4e does not learn a continuous embedded representation of a clinical chronology. Instead, it uses the un-embedded “bag-of-observations” vector described above as the encoding of the clinical snapshot.

**Phase II: Clinical Picture Inference.** Because the clinical snapshots provide incomplete information about the clinical picture of the patient (e.g., a snapshot produced from a radiology report describing a chest x-ray is unlikely to include many observations about other anatomical regions), the role of Phase II is to infer or impute the clinical picture of the patient as it relates to imminent pneumonia risk by combining and accumulating information from each clinical snapshot. We implemented Phase II by (1) casting the inferred clinical picture of the patient as the memory vector of a Recurrent Neural Network (RNN) and (2) training the RNN to infer what the clinical picture of the patient looks like after processing each snapshot and accounting for the elapsed time between it and the previous snapshot. Formally, for each encoded snapshot $d_t \in d_1, d_2, \ldots, d_L$ the RNN is trained to predict the clinical picture of the patient after elapsed time $\delta_{t+1}$ such that the final output (i.e., inferred clinical picture) computed from $d_L$ and prediction window $\delta$ captures sufficient information to predict the risk that the patient will develop pneumonia within $\delta$ days.

The RNN used for inferring the clinical picture is the core of PRONTO; consequently, the choice of memory unit used by the RNN is critical to the over-all success or failure of the model. We used a Recurrent Additive Network (RAN). RANs are a simplified alternative to LSTM- or GRU-based recurrent neural networks that use only additive connections between successive layers and have been shown to obtain similar performance with 38% fewer learnable parameters. The lower number of learnable parameters is ideally suited for deep learning with sparse datasets as it avoids the vanishing gradient problem and lowers the ability of the model to “memorize” the training set, improving generalizability. Formally, let $x_t \in m_1, \ldots, m_L$ represent the sequential input to the RNN such that $x_t := [d_t, \delta_{t+1}]$ and $x_L := [d_L, \delta]$, let $h_{t+1} \in [h_{L+1}, \ldots, h_L]$ represent the output of the RNN, and let $m_t \in m_1, \ldots, m_L$ represent the internal memory of the RNN after processing $x_t$. Figure 5 provides the equations used to compute $h_t$ with the RAN.

**Phase III: Pneumonia Prediction.** The final phase of PRONTO is to predict the imminent risk (i.e., probability) that the patient will develop pneumonia within prediction window $\delta$ using the inferred clinical picture $h_L$. We do this using a dense linear projection layer to produce a two-element vector followed by a softmax activation which computes the probability that the patient is at imminent/high- or low- risk for developing pneumonia within $\delta$ days:

$$ P\left( \text{pneumonia} \mid C, \hat{\delta} \right) = \text{softmax} \left( W_p h_L + b_p \right) $$

(1)

**C. Training PRONTO without Direct Supervision**

It is difficult to determine ground-truth labels for risk: risk is inherently difficult for humans to quantify, and using an existing metric of pneumonia risk to train the model would reduce the model to approximating the existing metric rather than learning to distinguish between low- and high-risk. Consequently, in this paper, we used indirect
supervision to train PRONTO. We first detected the onset of pneumonia within the chronology of each patient using the following criteria: (1) if any observation in a content section of the note was a descendent of pneumonia in the UMLS hierarchy; (2) if any observation in a content section of the note had the word pneumonia or any of its acronyms in its UMLS preferred name; and (3) if there was a non-negated mention of the word pneumonia or any of its acronyms in the content section of the note in a line that did not end with a colon. We defined the content sections of an ICU note as any section not corresponding to consults, family history, past medical history, or social history. Because some patients have long ICU stays in which the duration of pneumonia is not always clear, we then truncated our chronologies to end at the last snapshot before the onset of pneumonia. In this way, each chronology in the training set begins with one or more snapshots indicating the absence of pneumonia, \( O_1, \ldots, O_L \), followed by exactly one snapshot indicating the onset of pneumonia, \( O_{L+1} \). These chronologies were used as high-risk examples; i.e., \( P \left( \text{pneumonia} \mid C = \{O_1, \ldots, O_L; \delta_2, \ldots, \delta_L\}; \hat{\delta} = \delta_{L+1} \right) = 1 \). We created low-risk examples by randomly sampling a sub-sequence length \( S \in [1, L-1] \) and truncating each chronology to \( S \); i.e., \( P \left( \text{pneumonia} \mid C = \{O_1, \ldots, O_S; \delta_2, \ldots, \delta_S\}; \hat{\delta} = \delta_{S+1} \right) = 0 \). For example, the chronology illustrated in Figure 2 has the onset of pneumonia in the fifth snapshot and would be associated with two training examples: (1) the chronology including the first four snapshots would be considered as high-risk given the time window \( \hat{\delta} = 1 \) and (2) a randomly sampled chronology ending at the first, second, or third snapshot would be considered as low-risk, given the time windows \( \hat{\delta} \in \{1, 3, 4\} \), respectively. We trained the model by minimizing the cross-entropy loss using Adaptive Moment Estimation (ADAM) with the default initial learning rate \( \eta = 0.001 \).

**Dropout** The sparse and limited nature of our training data allows the model to avoid learning how to predict pneumonia risk and instead just memorize which chronologies are positive or negative for pneumonia based on rarely occurring observations or sequences of observations (i.e., over-fitting). To prevent over-fitting, we evaluated two forms of dropout: (1) layer dropout before each input to the RNN, and between the final output of the RNN and the softmax layer, and (2) vocabulary dropout in which random rows (e.g., observations) of the embedding matrix used in the embedding layer of each snapshot encoder was randomly set to a zero vector. Vocabulary-level dropout prevents the model from simply memorizing a rare subset of (confounding) clinical observations that happen to be associated with pneumonia in our dataset.

**D. Using PRONTO for Inference**

After training PRONTO to predict the disease-risk for a retrospective pneumonia cohort, it can be used to predict the pneumonia-risk for (new) patients in real-time. Specifically, for a (new) patient, we can predict the risk that he or she will develop pneumonia within a given prediction window \( \hat{\delta} \) by (1) extracting the clinical chronology from any ICU notes produced thus far in the patients ICU stay, and (2) using Equation (1) to predict the probability of pneumonia given \( C \) and \( \hat{\delta} \).

**Results**

To evaluate PRONTO, we used a test set of 146 at-risk patients as described in the Materials section. For each patient, we determined \( S \), the snapshot indicating pneumonia onset, and measured the performance of the model when correctly predicting (1) high-risk given the chronology proceeding \( S \) and the elapsed time between \( S \) and the chronology, and (2) low-risk for a random snapshot occurring in the chronology before \( S \). Specific details and examples are provided in subsection C of the Methods section. In this way, we produce a test set with an approximately 57% high-risk and 43% low-risk examples and are able to assess that the model can correctly predict when the risk of developing pneumonia progresses from low-risk to high- (imminent) risk.

We compared the performance of PRONTO against a support vector machine (SVM) and two simple baselines:

- **SVM** Inspired by Bejan et al. (2013),\(^{11}\) we trained a linear SVM to predict pneumonia risk using either (a) only the final snapshot of each training chronology, or (b) using all snapshots in each training chronology (where-in all but the final snapshot are associated with a low risk);
- **Constant** A simple baseline predicting all snapshots as (a) low risk, or (b) high risk; and
- **Random** A simple baseline predicting risk based on (a) the prior probability of high and low risk in the training chronologies, or (b) a uniform distribution over high and low risk.

All models were trained for twenty epochs, and the model parameters which resulted in the highest F\(_1\)-measure on the development set were used for testing.
We measured the Accuracy (A), Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, Precision (P), Recall (R), and F1-measure (the harmonic mean between Precision and Recall)\[^{14}\]. To investigate the susceptibility of each system to over-fitting, we report all five metrics on the training, development, and testing sets. Table 1 presents the performance of all baseline systems as well as the performance of PRONTO when using each of the clinical snapshot encoding architectures illustrated in Figure 4. To better understand the behavior of the PRONTO model, we also report the learning curve (measured using AUC) on the training, development, and testing data when using each type of clinical snapshot encoder in Figure 6.

Table 1: Performance when distinguishing between high/imminent and low pneumonia risk when using (1) Support Vector Machines, (2) constant baselines, (3) random baselines, and (4) PRONTO using different clinical snapshot encoding architectures.

<table>
<thead>
<tr>
<th>System</th>
<th>A</th>
<th>AUC</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>A</th>
<th>AUC</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>A</th>
<th>AUC</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM: Final Snapshot</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>1</td>
<td>0.98</td>
<td>0.58</td>
<td>0.53</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>0.60</td>
<td>0.54</td>
<td>0.72</td>
<td>0.69</td>
<td>0.70</td>
</tr>
<tr>
<td>SVM: All Snapshots</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.33</td>
<td>0.42</td>
<td>0.50</td>
<td>0.17</td>
<td>0.25</td>
<td>0.39</td>
<td>0.51</td>
<td>0.70</td>
<td>0.21</td>
<td>0.32</td>
</tr>
<tr>
<td>Constant: Low Risk</td>
<td>0.43</td>
<td>0.50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.44</td>
<td>0.50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.41</td>
<td>0.50</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Constant: High Risk</td>
<td>0.73</td>
<td>0.50</td>
<td>0.57</td>
<td>1</td>
<td>0.73</td>
<td>0.57</td>
<td>0.50</td>
<td>0.56</td>
<td>1</td>
<td>0.72</td>
<td>0.55</td>
<td>0.50</td>
<td>0.59</td>
<td>1</td>
<td>0.74</td>
</tr>
<tr>
<td>Random: Prior</td>
<td>0.54</td>
<td>0.49</td>
<td>0.66</td>
<td>0.65</td>
<td>0.66</td>
<td>0.55</td>
<td>0.51</td>
<td>0.67</td>
<td>0.64</td>
<td>0.66</td>
<td>0.54</td>
<td>0.47</td>
<td>0.67</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>Random: Uniform</td>
<td>0.49</td>
<td>0.49</td>
<td>0.56</td>
<td>0.49</td>
<td>0.52</td>
<td>0.48</td>
<td>0.49</td>
<td>0.56</td>
<td>0.43</td>
<td>0.48</td>
<td>0.44</td>
<td>0.44</td>
<td>0.51</td>
<td>0.41</td>
<td>0.45</td>
</tr>
<tr>
<td>PRONTO: RNN</td>
<td>0.65</td>
<td>0.64</td>
<td>0.67</td>
<td>0.76</td>
<td>0.71</td>
<td>0.63</td>
<td>0.63</td>
<td>0.69</td>
<td>0.62</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
<td>0.76</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>PRONTO: CNN</td>
<td>0.66</td>
<td>0.65</td>
<td>0.69</td>
<td>0.75</td>
<td>0.72</td>
<td>0.63</td>
<td>0.62</td>
<td>0.67</td>
<td>0.66</td>
<td>0.66</td>
<td>0.61</td>
<td>0.60</td>
<td>0.71</td>
<td>0.65</td>
<td>0.68</td>
</tr>
<tr>
<td>PRONTO: DAN</td>
<td>0.71</td>
<td>0.68</td>
<td>0.66</td>
<td>0.96</td>
<td>0.78</td>
<td>0.72</td>
<td>0.66</td>
<td>0.69</td>
<td>0.96</td>
<td>0.80</td>
<td>0.72</td>
<td>0.66</td>
<td>0.69</td>
<td>0.96</td>
<td>0.80</td>
</tr>
<tr>
<td>PRONTO: Dense</td>
<td>0.81</td>
<td>0.80</td>
<td>0.81</td>
<td>0.87</td>
<td>0.84</td>
<td>0.56</td>
<td>0.55</td>
<td>0.56</td>
<td>0.74</td>
<td>0.68</td>
<td>0.59</td>
<td>0.55</td>
<td>0.63</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>PRONTO: Sparse</td>
<td>0.70</td>
<td>0.68</td>
<td>0.70</td>
<td>0.81</td>
<td>0.75</td>
<td>0.60</td>
<td>0.58</td>
<td>0.62</td>
<td>0.72</td>
<td>0.67</td>
<td>0.64</td>
<td>0.61</td>
<td>0.70</td>
<td>0.74</td>
<td>0.72</td>
</tr>
</tbody>
</table>

\[^{14}\]Note: Due to the class imbalance, we emphasize the F1-measure in our study; the other metrics are reported to illustrate that, unlike the Constant: High Risk baseline, the learned approaches are able to also detect low risk, as evidence by Accuracy and AUC.

Implementation Details PRONTO was implemented in TensorFlow v1.12. Baseline systems were implemented using Sci-kit Learn\[^{25}\] v0.19.2. The width of observation embeddings, snapshot embeddings, and the internal memory used by the snapshot RNN were selected from \{100, 200, 300, 400, 500\} and determined to be 200; the number of kernels considered by the CNN was selected from \{100, 500, 1,000, 2,000\} and determined to be 1,000; the choice of k and j for the CNN and DAN encoders were selected from \{1,2,3\} and determined to be 2; the dropout and vocabulary dropout probabilities were selected between 0 and 1 in increments of 0.10 and determined to be 0.0 (no dropout) and 0.50, respectively. Vocabulary size was not limited, the maximum snapshot size was set to 256 (with an average of 141 observations per snapshot), and the maximum chronology length was set to 7 (with an average of 6 snapshots per visit). All baseline and PRONTO experiments used the same fixed random seed.

Discussion

As shown in Table 1 PRONTO using the Deep Averaging Network (DAN) for encoding clinical snapshots obtained the highest performance, exhibiting a 22% (relative) increase in AUC and a 12% (relative) increase in F1 performance on the testing cohort compared to the best performing SVM (SVM: Final Snapshot). The high performance of PRONTO, especially in terms of Recall, demonstrates that the observations reported in clinical notes alone can be predictive of imminent pneumonia risk. However, the poorer performance of the SVM baselines when using the same observations suggests that deep learning methods can more effectively harness clinical texts even with small datasets. Our results also indicate that model design – particularly in small data settings – has a substantial impact on performance, with sub-optimal models performing similarly to an SVM (in terms of F1). When comparing SVM: All Snapshots to SVM: Final Snapshot, we can see the importance of balancing the distribution of high- and low-risk examples when training the model, indicating that even when the dataset is limited, omitting potential training data can lead to improved generalizability. Interestingly, the dense representation of clinical snapshots provided the highest performance on the training data, but performed poorly compared to other methods on the development and testing set. The DAN, Dense, and Sparse methods viewed observations as unordered set, indicating that the more complex sequential representations may be too expressive to generalize from sparse data.

When examining the learning curves in Figure 6 we can see that all approaches but the DAN quickly over-fit the training
data, suggesting that DANs are better able to generalize when used on sparse and limited datasets. Interestingly, despite being the most complex snapshot encoder we considered, the RNN had the lowest performance, suggesting that considering the sequential order of clinical observations within a clinical snapshot had little impact. We believe this is due to the fact that by reducing clinical notes to observations we are removing much of the context around each observation and likely making their order insignificant.

Finally, we examined the accuracy of the disease onset labels we automatically produced and used for indirect supervision. Specifically, we selected difficult-to-label notes (i.e., notes without pneumonia which indicated infiltrates and/or fever and notes with pneumonia without any mention of infiltrates and/or fever) and randomly selected 50 such notes for manual judgment by two physicians. The two physicians had moderate agreement (82% simple agreement; Cohen’s $\kappa = 0.49$). We first measured the accuracy of disease onset labels by considering a label as correct if it agreed with either physician, with an accuracy of 80%. We also determined an accuracy of 78.6% when measuring the accuracy of labels only for notes with agreement between both physicians. These results suggest that automatic detection of pneumonia onset with natural language processing is fairly reliable, and well suited for indirect supervision.

A. Limitations and Error Analysis

There are a number of limitations to this study beyond the limited dataset. First, we used automatic natural language processing to identify clinical observations from ICU notes. While UMLS allows some level of normalization, we observed cases in which the clinical observations were underspecified. For example, we found observations such as “aim”, “probe”, and “distillation” which are ambiguous without context. Second, we found that the attributes detected by ConText were not always reliable, introducing noise into the clinical observations and, indirectly, into the labels used to train the model. In future work we plan to qualify observations by their contextual attributes (experiencer, negation, etc.) to provide more nuanced information to the model. Third, our model considers each ICU stay as an independent chronology. It is possible that information from a previous admission could impact the risk of a patient developing pneumonia (i.e., by allowing the model to distinguish between new observations and re-occurring observations). Fourth, we used an approximately equal number of negative and positive examples when evaluating the model. While this was an intentional decision to make it easier to compare the relative strengths and weaknesses between different models, it does not give an accurate indication of clinical performance where-in one would expect substantially more low-risk patients. Finally, in this initial study we examined only pneumonia. In future work, we would like to predict other types of health-care associated infections, evaluate in a more clinically-realistic setting, and explore the inclusion of charted information.

Conclusion

In this paper, we described a data-driven deep learning model for distinguishing between low- and high-risk of HAP in an at-risk ICU patient population. We show that not only can our model be successfully trained on incomplete, limited, and noisy data – obtaining 96% Recall, 72% AUC, and 80% $F_1$-measure with less than 1,500 examples – but that it outperforms SVM-based models using the same features by 22% (AUC). Moreover, we empirically evaluated the impact of five different strategies for encoding clinical observations, determining that Deep Average Networks provide the most reliable encoding of multiple clinical observations. We believe our results demonstrate that ICU notes provide sufficient information to predict and distinguish between high and low pneumonia risk within an arbitrary, given time window without feature engineering or structured data. It is our hope that our findings enable deep learning to be more easily applied in other scenarios with incomplete or limited training data.
Reproducibility
The source code for PRONTO is available on GitHub at https://github.com/h4ste/pronto

Acknowledgements
This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References
Exploring Different Approaches in Measuring EHR-based Adherence to Best Practice – A Case Study with Order Sets and Associated Outcomes

Nathan C. Hulse, PhD,1,2, Jaehoon Lee, PhD,1,2 José Benuzillo, MS

1Intermountain Healthcare, Salt Lake City, UT; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT;

Abstract

In connection with a recent enterprise-wide rollout of a new electronic health record, Intermountain Healthcare is investing significant effort in building a central library of best-practice order sets. These order sets represent best practice guidelines for specific clinical scenarios and are deployed with the intent of standardizing care, reducing variation, and consistently delivering good clinical outcomes to the populations we serve. The importance of measuring their use and the level to which caregivers adhere to these standards becomes an important factor in understanding and characterizing the impact that they deliver. Notwithstanding the importance of these metrics, well-defined methods for measuring adherence to a given clinical guideline as delivered through an order set are not fully characterized in the medical literature. In this paper, we describe initial efforts at measuring compliance to a defined ‘best practice’ standard by means of content utilization analysis, a calculated adherence model, and relevant clinical key performance indicators. The degree to which specified clinical outcomes vary across these measurement models are compared for a group of order sets tied to treating coronary artery bypass graft patients and heart failure patients. While the patterns derived from this analysis show some uncertainty, more granular methods that look at line-item, or ‘order level’ detail reveal more significant differences in the corresponding set of outcomes than higher-level adherence surrogates.

Introduction

Order set libraries have become an established part of electronic health record (EHR) implementation as EHRs have become increasingly common in the United States. Previous research in the informatics field has shown them to be convenient, thought-provoking, and foundationally supportive of clinical best practice. Studies have shown that well-crafted order set libraries are among the key determinants that predict a successful deployment of a new EHR.

Order sets are a unique component of an EHR installation in that (like clinical decision support rules) they represent clinical knowledge as delivered through interactive user interfaces. They are typically rendered as interactive forms that deliver logical, pre-populated views of commonly grouped orders for specific procedures or clinical conditions. Their overall intent is to help clinicians deliver appropriate care in a simple, streamlined way. (Figure 1 below depicts an order set in the EHR used at Intermountain). The ‘checklist effect’ that they deliver forces users to mentally process and remember steps which could otherwise be easy to overlook in clinical care. Collectively, as caregivers routinely use standardized order sets in delivering care, several key objectives that health organizations hope to realize would include lower overall variation in clinical care processes and by correlation, improved clinical outcomes and lower associated costs.

The benefit afforded by these order set content libraries comes with an associated need for ongoing investment in resources so as to maintain and update the content to ensure validity against current clinical science. Both homegrown and vended order set libraries have arisen in recent years as implementers have noted and characterized the difficulties of building up and maintaining a comprehensive library of order sets. Compiling and maintaining an enterprise knowledge base of order sets has been described as highly resource-intensive. The magnitude of work required to maintain the content contained in these order set knowledge bases has led various groups to implement various automated and machine-learning-based processes intended streamline the review and update workflows that support them. Researchers have noted that with the exponential growth in medical knowledge, that there is a real risk that once implemented, order sets may be inadequately maintained; in essence, driving caregivers to practice outdated medicine on a widespread basis.

Ultimately, the variability of functional pathways afforded by an EHR often allows caregivers multiple ways of performing the same task. Local experience has shown that clinical tools like quick-lists, generic order set templates (e.g. ‘general surgery order set’) personalized derivatives of order set templates (sometimes known as ‘favorites’ or ‘personal order sets’) and even poorly-coordinated central authoring efforts can leave users with a content library that affords them different content vehicles targeted at the very same clinical process. Furthermore, order set usage is not typically mandated by process or workflow, and users are free to electronically order care patterns without them.
entirely. Implementers must be aware of and confront the need to standardize the way that the care teams they manage use centralized order sets through education, active monitoring, and ongoing feedback. Figure 2 illustrates how treatment of a cardiovascular care processes (with acute myocardial infarction depicted as a relevant example) has a broad set of different content utilization patterns as it relates to order set content was used as part of the care delivery process. In this figure, we depict order sets which are designated as an intended standard of care for this condition in the striped bars, whereas the solid bars indicate other content templates that were used. The distribution shows a wide variety of order set content that is used in caring for this single condition. Some of this diversity is to be expected, in the treatment of comorbidities and in addressing personalized needs. Yet, the chart shows the diversity of knowledge content that can address a common condition, and the associated difficulty of standardizing care. In short, publishing thoughtful and relevant content standards as order sets does not ensure they will be uniformly embraced.

<table>
<thead>
<tr>
<th>Component</th>
<th>Status</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV Acute Coronary Syndrome Probable Phased, PSO - Patient Status Order (Planned Pending)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admit/Transfer/Discharge/Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non Categorized</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Powerplan Indicator</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV Acute Coronary Syndrome Probable Phased, Orders (Planned Pending)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Care</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vital Signs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiac Monitoring*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intake and Output*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lines/ Tubes/ Drains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peripheral IV Insert</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bedrest with Bathroom Privileges</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diet/ Nutrition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Except for Medications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTE: If a medication is not available for the facility of the active encounter please contact pharmacy for alternate or to request stock</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitroglycerin (Nitrostat)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4 mg, Sublingual, every 5 min, PRN...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspirin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>324 mg, Oral, Once, Tab-Chew Hold if given in the ED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heparin Acute Coronary Syndrome (ACS) Limited</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.5 mg, Oral, every 6 hr, Tab Hold if Heart Rate is less than 60 BPM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 - Order set example. This particular order set focuses on treatment of acute myocardial infarction

© Cerner Corporation. All rights reserved. This document contains Cerner confidential and/or proprietary information belonging to Cerner Corporation and/or its related affiliates which may not be reproduced or transmitted in any form or by any means without the express written consent

It is important to note that variation in order set usage can occur at two different levels. First, caregivers still have the ability to use different content in delivering care (e.g. use a designated clinical standard or something entirely different), but also have the ability to vary in terms of how closely they follow the care outlined within an enterprise standard itself. Order set templates are inherently interactive, and users have the ability to select orders that aren’t part of the template, remove recommended orders, or change default values and order sentences. Inevitably there will be differences not only in what content is or isn’t used, but also in how users interact with the content itself. Measurement models aimed at measuring adherence to a standard of care must be able to account for differences at multiple levels of engagement.

Background

Intermountain Healthcare is a not-for-profit integrated delivery network that serves the populations of the Intermountain West (Utah and southern Idaho). It has 22 hospitals, over 150 clinics, a medical group of over 700 employed physicians and an insurance plan that serves the needs of the people in the region. Collectively, it accounts
for roughly half of the healthcare given in the region and insures a little more than a quarter of the population. In switching EHRs and the toolsets that support them, our main priorities in terms of order set knowledge management were threefold:

1. Build tools and processes to quickly amass and refine a sufficiently comprehensive knowledge base,
2. Provide visibility into usage patterns surrounding their use, including actionable recommendations for improvement
3. Characterize clinical caregivers’ adherence to these standards and their associated impact in clinical outcomes.

Previous efforts have detailed efforts in deriving actionable recommendations for change to order set templates based off collective usage data\(^8,10,15\). In this manuscript, we will focus describe current efforts at characterizing how well a clinical standard is used and the associated clinical outcomes derived from this care.

**Rationale**

We want to be able to clearly understand the implications of owning and publishing an order set library. In order to fully understand the value of our efforts, we want to understand both the extent to which users are engaging with a
clinical content standard (e.g. centrally defined and designate order set), but how they are using it and the combined effects of the above as it relates to clinical outcomes. Our hypothesis in pursuing this investigation is that the more that our measurement algorithms accounts for specific actions, the greater will be our ability to account for any observed differences in the corresponding outcomes. We intend to account for three specific approaches to measurement of adherence in this investigation:

1. Whether or not a designated enterprise content standard was used as part of an encounter for which it was designed
2. A calculated approach that quantifies how closely a user that engaged with the standard
3. A ‘key performance indicator’ approach in which specific key electronic orders are accounted for, regardless of whether the orders originated from a designated enterprise standard order set

Methods

In order to facilitate the study, we engaged with colleagues in the cardiovascular clinical program at Intermountain Healthcare to identify two specific diseases of interest, specifically acute myocardial infarction and heart failure. After discussion with them, we determined to look at ordering patterns related to standards of care intended for treating patients presenting with those conditions. For each, we agreed upon a common cohort definition for the intended targets of specific enterprise content sets. We also assembled a common set of outcomes of interest related to the care processes decided upon. Eight order sets related to treating acute myocardial infarction and six order sets relevant in treating heart failure were identified. These cohort definitions and order sets of interest are detailed in Table 1.

Table 1- Cohort definitions and enterprise standard order sets for acute myocardial infarction and heart failure

<table>
<thead>
<tr>
<th>Cohort Definition</th>
<th>Acute Myocardial Infarction</th>
<th>Heart Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated</td>
<td>Patients age (Admission Date minus Birthdate) greater than or equal to 18 years admitted to the hospital for inpatient acute care with an ICD-9/10-CM Principal Diagnosis Code for AMI as defined by CMS.</td>
<td>Patients age (Admission Date minus Birthdate) greater than or equal to 18 years admitted to the hospital for inpatient acute care with an ICD-9/10-CM Principal Diagnosis Code for HF as defined by CMS.</td>
</tr>
<tr>
<td>Enterprise Order Sets to be Analyzed in the Study (each are designated standards for the associated conditions)</td>
<td>• CV Acute Coronary Syndrome Probable Phased&lt;br&gt;• CV Cath / PCI FLOOR to Cath Lab PRE Procedure Phased&lt;br&gt;• CV Cath / PCI POST - Recovery to Acute Care Floor Phased&lt;br&gt;• CV Cath / PCI POST - Recovery to Home Phased&lt;br&gt;• CV Cath / PCI POST - Recovery to ICU Phased&lt;br&gt;• CV Cath / PCI PRE Procedure Phased&lt;br&gt;• CV General Floor Admission Phased&lt;br&gt;• CV General ICU Admission</td>
<td>• CV Heart Failure Admission&lt;br&gt;• CV Heart Failure Discharge Phased&lt;br&gt;• CV Heart Failure ICU Admission&lt;br&gt;• CV Heart Failure Observation&lt;br&gt;• CV General Floor Admission Phased&lt;br&gt;• CV General ICU Admission</td>
</tr>
</tbody>
</table>

480
Adherence Measurement Definition 1 – Raw Usage Data - In discussing options surrounding various ways of measuring adherence to a clinical standard, we felt that the act of ordering from a designated order set during an encounter noted as relevant to the content. Simply put, this type of measurement would answer the question of whether or not the designated standard content was regularly used in the contexts for which it was designed. While this type of measurement doesn’t account for the specific actions that occur in a caregiver’s interactions with the order set form itself, it may serve as a sufficiently granular approximation of compliance to the standard since the effort to use it may in and of itself signal an intent to follow the pattern.

Adherence Measurement Definition 2- Calculated adherence scores, based off interactions within order sets – A second approach that we opted to pursue in measuring compliance involves a deeper measurement of how closely the issued orders resemble those that are predefined by the order set itself as a recommendation. In this model, users are credited for issuing orders that are recommended (as defined by preselection in the order set template), credited to a lesser extent for issuing optional orders in the order set template, and penalized for omitting recommended orders and adding ‘ad hoc’ orders to the ordering session (orders that are not included in the template at all). In building these score guidelines, we recognize that caregivers will have the need to deviate from a specific standard and tailor the care that they give. In attempting to assign scores and numbers to these types of interactions, we are attempting only to measure the overall adherence or deviation from the initial order set template, and not offer any critique or commentary as to the clinical appropriateness of the actions taken. In this model, the coefficients for these scoring mechanisms per line item order are detailed in the scenarios below:

- A = Order pre-selected in template & ordered = 1
- B = Order pre-selected in template & not ordered = -.5
- C = Order initially unselected in template, but ordered = 0.1
- D = Order initially unselected in template, and not ordered = 0.0
- E = Ad-hoc order not derived from order set template = -.5

In order to facilitate comparison of calculated adherence scores across order set groupings of variable size, we normalized each of these scores by dividing this score by the total number of orders issued in that encounter. This would allow scores to be meaningful and comparable from scores derived both from small and large templates. We pursued two normalization approaches, one accounting for a maximum score based off the total number of prechecked orders in the template, the second based off the total number of orders placed. The equations below where the variables are measures of the scenarios from the bullet list above were calculated for each ordering scenario for both heart failure and acute myocardial infarction.

\[
\text{Normalized score 1} = \frac{(A) + 0.1(C) - 0.5(B) - 0.5(E)}{\text{Total # of prechecked orders in template}}
\]

\[
\text{Normalized score 2} = \frac{(A) + 0.1(C) - 0.5(B) - 0.5(E)}{(A+C+E)}
\]

Adherence Measurement Definition 3- Key Performance Indicators- In this third adherence definition, we looked to simplify adherence metrics somewhat by focusing strictly on a ‘Key Performance Indicator’ (KPI) in which all predefined orders from an order set template were listed as KPI’s. We opted to pursue this approach rather than a ‘clinical expert defined’ set as we are attempting to explore generalizable methods of measuring adherence that do not necessitate extra steps of human involvement. We interpreted the order set authors’ intent in pre-selecting an order in the template as a stronger indicator that the associated order should probably be issued for the scenario. We then calculated a total count of orders that matched these lists in the encounters that matched the acute myocardial infarction and heart failure scenarios, regardless of whether these orders were derived from the enterprise content standards or not. To normalize these scores, we calculated the total number of KPI’s ordered in the encounter and normalized it by dividing by the total number of possible KPI’s tied to that scenario.

We extracted 12 months of order set usage data for the order sets detailed in Table 1 from Intermountain Healthcare’s enterprise data warehouse, deriving data only from our Cerner EHR system. Data was cleansed and vetted to remove test patients and other non-clinical data sources. The dataset used in the study was de-identified and vetted to ensure that PHI was not available. After extracting relevant ordering data from the encounters identified, we also pulled outcomes data from the cardiovascular clinical program’s outcomes database to align the order instances and scores with specific outcomes data.
In aligning data to specific outcomes results, we opted to group the data in patterns that match the three adherence measurement approaches detailed above. For Adherence Measurement Definition 1, we grouped all providers who used the enterprise standards in one cluster and those who didn’t in another. Summary statistics were prepared for each group. For Adherence Measurement Definition 2, we took the ‘adherent’ group from definition 1, and clustered it into three groups, based off the adherence scores tied to each encounter. Encounters were grouped by score into the above 75th percentile, 25th-75th percentile, and below 25th percentile groups. In similar fashion, Adherence Measurement Definition group 3, we grouped all participants into percentile-based groups. An important difference in this metric compared to the second group is that all participants were measured the same way, since it was not a pre-requisite to use the order set standard at all in order to still measure KPI scores.

**Results**

According to the calculation patterns and data extraction efforts above, we gathered data and summarized basic statistics for each measurement pattern. 1,693 instances of acute myocardial infarction and 1,736 instances of heart failure were identified from the cohort definitions and the 12 month data set from the Cerner EHR system. Table 2 contains the results for Adherence Measurement Definition 1, in which encounters were grouped according to whether or not the ordering provider used the designated order set standard for the specific encounter.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acute Myocardial Infarction</strong></td>
<td><strong>Heart Failure</strong></td>
<td><strong>Acute Myocardial Infarction</strong></td>
<td><strong>Heart Failure</strong></td>
</tr>
<tr>
<td>Avg Length of Stay</td>
<td>3.585 days</td>
<td>3.342 days</td>
<td>4.51 days</td>
</tr>
<tr>
<td>Avg Mortality rate</td>
<td>0.044</td>
<td>0.0485</td>
<td>0.0878</td>
</tr>
<tr>
<td>Avg Readmission Rate (30 days)</td>
<td>0.060</td>
<td>0.063</td>
<td>0.170</td>
</tr>
</tbody>
</table>

For Adherence Measurement Definition 2, we produced similar statistics, as shown in Tables 3 and 4. Histograms representing the distribution of scores in this exercise are shown in Figure 3 below. These distribution curves are mostly Gaussian in nature, though there are elements that might suggest a bi-modal normal distribution, implying that there may be reason to suspect that the data represents two heterogenous groups of data sources. Although we did not analyze further of what distinguishes the two groups in this study, we feel that it is statistically reasonable to group by <25, 25-75 and >75 percentiles from these bell-shaped distributions. In these clusters, we created three groups for the populations of caregivers that adhered to the enterprise standard order sets. After the scores had been calculated and normalized against the two denominator definitions above, we grouped the scores into three clusters, those above the 75th percentile, those between 25th and 75th, and the bottom quartile (below 25th percentile).

**Table 2-** Summary statistics for outcomes tied to encounters for acute myocardial infarction and heart failure, grouped by whether or not the ordering provider used the designated order set standard for the specific encounter.

**Figure 3 -** Histograms depicting the distribution curves of scores derived for AMI (left) and MI (right) data sets. The data is mostly normal, Gaussian in nature, though there are elements of bimodality in the sample.
Table 3- Summary statistics for outcomes tied to encounters for acute myocardial infarction, grouped by the scoring mechanism detailed in Adherence Measurement Definition 2. Results normalized against both denominator definitions are given.

<table>
<thead>
<tr>
<th>Acute Myocardial Infarction – Normative Denominator 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y - &gt;75th percentile</td>
</tr>
<tr>
<td>Avg Length of Stay</td>
</tr>
<tr>
<td>Avg Mortality rate</td>
</tr>
<tr>
<td>Avg Readmission Rate (30 days)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acute Myocardial Infarction – Normative Denominator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;75th percentile</td>
</tr>
<tr>
<td>Avg Length of Stay</td>
</tr>
<tr>
<td>Avg Mortality rate</td>
</tr>
<tr>
<td>Avg Readmission Rate (30 days)</td>
</tr>
</tbody>
</table>

Table 4- Summary statistics for outcomes tied to encounters for heart failure, grouped by the scoring mechanism detailed in Adherence Measurement Definition 2. Results normalized against both denominator definitions are given.

<table>
<thead>
<tr>
<th>Heart Failure – Normative Denominator 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;75th percentile</td>
</tr>
<tr>
<td>Avg Length of Stay</td>
</tr>
<tr>
<td>Avg Mortality rate</td>
</tr>
<tr>
<td>Avg Readmission Rate (30 days)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Heart Failure – Normative Denominator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;75th percentile</td>
</tr>
<tr>
<td>Avg Length of Stay</td>
</tr>
<tr>
<td>Avg Mortality rate</td>
</tr>
<tr>
<td>Avg Readmission Rate (30 days)</td>
</tr>
</tbody>
</table>

For Adherence Measurement Definition 3, we produced summary statistics, as shown in Tables 5 and 6. In these clusters, we created three groups for the populations, following similar percentile-based clusters as defined above. In this scenario, percentiles were based off KPI definition scores as previously detailed in the methods section.

Table 5- Summary statistics for outcomes tied to encounters for acute myocardial infarction, grouped by the scoring mechanism detailed in Adherence Measurement Definition 3.

<table>
<thead>
<tr>
<th>Acute Myocardial Infarction</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;75th percentile</td>
</tr>
<tr>
<td>Avg Length of Stay</td>
</tr>
<tr>
<td>Avg Mortality rate</td>
</tr>
<tr>
<td>Avg Readmission Rate (30 days)</td>
</tr>
</tbody>
</table>

Table 6- Summary statistics for outcomes tied to encounters for heart failure, grouped by the scoring mechanism detailed in Adherence Measurement Definition 3.

<table>
<thead>
<tr>
<th>Heart Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;75th percentile</td>
</tr>
<tr>
<td>Avg Length of Stay</td>
</tr>
<tr>
<td>Avg Mortality rate</td>
</tr>
<tr>
<td>Avg Readmission Rate (30 days)</td>
</tr>
</tbody>
</table>
Discussion

In exploring the various approaches to measuring adherence to best practice standards through the EHR, we found various patterns in terms of how outcomes differ, based off the generalized approach to measurement. The first approximation to adherence, specifically that of whether or not a designated content standard was used, showed relatively minimal differences in the outcomes distributions that corresponded with each group. Across the two diseases studied, the six outcomes metrics were all comparable, with three of the ‘content standard user group’ content producing slightly more favorable outcomes, and three producing slightly less favorable ones. Overall, whether or not the designated content sets were used did not offer strong differences in the outcomes analyzed in this effort.

Adherence measurement two took a slightly more detailed approach at analyzing the specifics of what users were doing with content inside the order sets of interest. For the acute myocardial infarction data set, almost all of the outcome metrics showed improvements in the highest quartile of adherence measures, with the exception of readmission rate under the second normative denominator. The heart failure data results exhibit similar trending that tends to correlate with stratification of outcomes favoring better outcomes with closer adherence, with the exceptions of length of stay.

The third approach to adherence, specifically that of only analyzing pre-checked orderables and treating them as KPIs, showed a trend for all three outcomes trending to better outcomes as ‘adherence’ to the KPI’s increased for the acute myocardial infarction data. The heart failure showed similar trends in comparing the bottom and middle quartiles, but less so in the highest quartile. In conferring with clinical domain experts on this observation, they noted that the diagnosis and treatment of heart failure is inherently broader by nature and the underlying etiology and associated treatments are more varied.

The approach to normalizing the data scores in the second and third branches of our study seemed to allow for comparative analysis of content sets of varying sizes. By nature, labeling ‘groups’ of order sets as a single standard makes comparison a much more difficult task. A similar type of study in which single order sets were designated as a content standard would make for a much simpler comparison.

One of the underlying premises of our effort is that we would hope and expect that as adherence to clinical standards was measurably improved, that corresponding outcomes would trend in a positive direction. The data we have analyzed doesn’t hold this as universally true, though there are many other co-variates not analyzed by our current approach. Our models are ‘order-centric’ and don’t account for co-morbidities, actual delivery of care, patient compliance, and other factors that have an important bearing on outcomes.

Limitations

We are only basing our preliminary analysis off two content types from a single clinical knowledge domain in cardiovascular medicine. Whether or not similar analyses would be relevant for other clinical domains remains to be seen. Our current model does not account for acuity, and that alone may account for significant impact in terms of how the corresponding outcomes move. Further, this analysis only looks at data from one site and one EHR vendor.

In this analysis, we did not assume any variation existed in how the providers used order sets. We intentionally did not use data derived from order set from use at early stages of implementation so that we could avoid inadvertently incorporating some learning curve effects in the data. However, there are many studies that show that user behaviors in interacting with order sets differ by physician characteristics, clinical specialty, clinical settings, and familiarity or training level to computerized order entry systems16. A more rigorous statistical effort to show that those factors do not cause variations of ordering patterns may be required, as a preliminary assumption to analyze the relations between the ordering patterns and clinical outcomes. To do so, further order set utilization data may be required to analyze it at different levels of granularity.

Next Steps

We intend to pursue similar analyses in other enterprise content standards, including surgical targets of appendectomy, cholecystectomy, and joint replacement surgery. We are hoping to see if similar patterns hold or if more defined ones emerge. Some of these content standards involve a single order set instead of groups of order sets as a designated standard, so we plan to test out similar methods that do not require a normalization step so as to be able to compare scores. Finally, we also intend to test machine-learning models that can derive and identify KPI types of order items
that may be justifiably more predictive and potent than just the set of preselected orders in the original order set template.

**Acknowledgments**

We would like to acknowledge our colleagues at Cerner for their help in identifying data relevant to the order set templates, corresponding metadata, and the order set instance derivatives created from the use of these plans in the clinical data repository. Additionally, we would like to thank clinical data analysts and domain experts from the Cardiovascular Clinical Program at Intermountain Healthcare for their time, feedback, and suggestions on the approaches taken in the manuscript.

**Conclusion**

In this paper, we have characterized the need for methods to calculate adherence to clinical standards, including artifacts distributed through knowledge artifacts like order sets. Three separate approaches for calculating adherence have been described and characterized, including high level content usage, a more detailed mathematical approach for approximating adherence, and a third approach that focuses only on key actions. The various scores and corresponding outcomes for all of these approaches have been presented for two key disease processes of interest, specifically for acute myocardial infarction and heart failure. We hope to extend our models going forward to more fully account for user behaviors in adhering to and varying from clinical standards, as well as the corresponding effects that accompany them in clinical care.

**References**

Trust and Privacy: How Patient Trust in Providers is Related to Privacy Behaviors and Attitudes

Bradley E. Iott, MPH1, Celeste Campos-Castillo, PhD2, Denise L. Anthony, PhD1
1University of Michigan, Ann Arbor, Michigan, USA; 2University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Abstract

Healthcare today requires extensive sharing and access to patient health information. The use of health information technology (health IT) exacerbates patients’ privacy concerns because it expands the availability of patient data to numerous members of the healthcare team. Patient concerns about the privacy of their data may be associated with nondisclosure of their information to providers. Patient trust in physicians, a multi-dimensional perception influenced by patient, physician, and situational factors, can facilitate disclosure and use of health IT. Previous work has done little to explore how specific dimensions of trust in physicians are related to patient information-sharing concerns or behavior. Using data from a nationally-representative survey, we show that patients with higher trust in provider confidentiality have significantly lower likelihood of reporting having ever withheld important health information and lower likelihood of thinking it is important to find out who has looked at their medical records. Patient trust in physician competence is related to higher likelihood of thinking it is important for health care providers to share information electronically. This work sheds light on the importance of considering multiple dimensions of trust for patient behavior and attitudes related to their information sharing with health care providers.

Introduction

The delivery of healthcare today requires extensive sharing and access to patient health information. The use of health information technology (IT) expands the availability of patient data to numerous healthcare providers throughout and between healthcare organizations. Since the widespread expansion of electronic health record (EHR) implementation following the distribution of incentives by the HITECH act, use of Internet-enabled applications for health (e.g., patient portals, mHealth apps) has increased steadily over the past decade. Use of health IT has yielded numerous positive benefits for patients, including increased access to data and improved efficiency1. However, such benefits come with increased concern about the potential for privacy breaches2,3. Indeed patient concerns about the privacy of their health information can impede their access to health care and hinder disclosure to providers4-7, creating incomplete medical records. Complete information sharing by patients with providers is necessary to ensure the quality and accuracy of data available in health information systems.

Patient trust in physicians, a multi-dimensional perception8-10 can potentially remedy the issue. Patient trust in physicians is influenced by patient, physician, and situational factors, and can facilitate disclosure9 and use of health IT11. One dimension, in particular, seems to matter more than others for ameliorating privacy concerns6,12: trust in physician confidentiality. However, few quantitative studies of trust in physicians explore multiple dimensions and those that do have only single measures of each dimension, leaving critical gaps in knowledge about how trust may matter8-10.

Systematic disparities in patient trust in physicians also exist, though again the dimensions of trust matter differently for different groups. Some research finds that racial and ethnic minorities have lower levels of certain dimensions of trust in physicians,13,14 and other research shows that patients with devalued health characteristics have lower trust in physician confidentiality specifically12. Similarly, disparities in use of health IT exist between racial groups and between those with stigmatized health conditions, suggesting the need to further investigate how different dimensions of trust may be related to patient attitudes and behavior related to information sharing.

Here we analyze national survey data to determine how dimensions of patient trust in physicians are related to patient disclosure of health information and to attitudes about information sharing within health care (see Figure 1). Figure 1 situates our analysis in the context of existing work. Our analysis specifically focuses on the association between three dimensions of trust in physicians and patients’ attitudes towards privacy and health information disclosure behavior (represented by the red arrow in Figure 1). Associations between patient demographic characteristics, health status, trust in physician, and health information attitudes and behaviors are indicated by blue directional arrows. Finally, the association between accuracy and quality of data in health IT systems and health
information attitudes and behaviors is represented by a dashed line, as this is an indirect consequence of patients’ disclosure behaviors and a needed focus of future work.

**Figure 1.** Study Framework: Situating our Analysis

*The red arrow represents the focus of our analysis.*

**Methods**

We conducted a cross-sectional, nationally-representative mail survey in 2014 to five thousand households across the United States based on three separate samples of households. For the general population sample, 1800 households were selected. For the Hispanic oversample, 1600 households were selected from the five communities with the highest concentration of Hispanics in the US: East Los Angeles, CA; Laredo, TX; Hialeah, FL; Brownsville, TX; and McAllen, TX. The African-American oversample included 1600 households from the following communities: Detroit, MI; Jackson, MS; Miami Gardens, FL; Birmingham, AL; and Baltimore, MD. In total, 5,000 households were selected. A total of 784 households completed the survey. The response rate of the survey was 15.7%. However, from the original sample of 5,000 households, 481 surveys were returned as undeliverable with no forwarding address available. The survey was conducted by the University of Nebraska Bureau of Sociological Research in partnership with Dartmouth College and was funded by HHS 90TR0003/01.

Data weights account for sample design (including the three sampling frames), nonresponse, and population characteristics to be representative of the national population.

Respondents were asked whether they had ever withheld information from their provider because of privacy and security concerns (1=Yes/0=No). Respondents were also asked two questions about attitudes towards privacy, namely, how strongly they agreed that it is important 1) to find out who has looked at their medical record and 2) that providers share their health information electronically (for each, strongly agree=1/less than strongly agree=0).

The survey also assessed several dimensions of trust in physicians, based on previous literature: trust in physician confidentiality, trust in physician honesty, trust in physician competence, and overall trust in physician. Each dimension was measured with items using a 5-point Likert scale. Trust in physician confidentiality is an aggregate measure created as the average response to three statements, namely “I trust my doctor to keep personally sensitive information private,” “I worry that private information my doctor has about me could be used against me,” and “I worry that my doctor may share embarrassing information about me with people who have no business knowing it.” Trust in physician honesty is an aggregate measure created as the average response to two statements, “I feel my doctor does not do everything s/he should for my medical care” and “I trust my doctor to perform only medically necessary tests and procedures.” Trust in physician competence is an aggregate measure created as the average response to three statements, “I trust my doctor to offer me high quality care,” “my doctor is a real expert
in taking care of medical problems like mine\textsuperscript{15,16}, and “my doctor’s medical skills are not as good as they could be” (reverse coded)\textsuperscript{16,17}. An overall measure of complete trust is included as the response to the statement “all in all, I have complete trust in my doctor\textsuperscript{16,17,19,20}.”

Data analysis was conducted using Stata 15.1. Survey data was analyzed using population sample weights. Logistic regression models were fitted to predict the three binary categorical dependent variables: having ever withheld information from providers due to concern about privacy, strongly agreeing that it is important that healthcare providers share my information electronically, and strongly agreeing that it is important that I can find out who has looked at my medical records. Independent variables are the four measures of trust: trust in confidentiality, honesty, competence, and overall trust. Each model controls for respondent demographic characteristics, self-rated health, health insurance status, and care quality (see Table 1 for descriptive statistics).

**Results**

Table 1 provides unweighted demographic details about the study sample. The majority of respondents in the sample are white (84.6%), educated with at least a BA/BS degree (58%), have excellent or good self-rated health (84.9%), have health insurance (96.0%), and report receiving very good or excellent health care in the previous year (60.5%).

**Table 1.** Descriptive statistics of Consumer Characteristics, Analytic Sample N=542, unweighted

<table>
<thead>
<tr>
<th>Variables</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>51.6</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
</tr>
<tr>
<td>African American, non-Hispanic</td>
<td>8.8</td>
</tr>
<tr>
<td>Hispanic</td>
<td>6.6</td>
</tr>
<tr>
<td>White, non-Hispanic</td>
<td>84.6</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td>18 – 29 years of age</td>
<td>8.7</td>
</tr>
<tr>
<td>30 – 39 years of age</td>
<td>20.5</td>
</tr>
<tr>
<td>40 – 49 years of age</td>
<td>21.9</td>
</tr>
<tr>
<td>50 – 59 years of age</td>
<td>22.0</td>
</tr>
<tr>
<td>60+ years of age</td>
<td>26.8</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
</tr>
<tr>
<td>High School or less</td>
<td>15.9</td>
</tr>
<tr>
<td>Some College</td>
<td>26.1</td>
</tr>
<tr>
<td>BA/BS College Degree</td>
<td>35.7</td>
</tr>
<tr>
<td>Graduate Degree</td>
<td>22.3</td>
</tr>
<tr>
<td><strong>Excellent or Good Self-rated Health</strong></td>
<td>84.9 (0.02)</td>
</tr>
<tr>
<td>Has health insurance</td>
<td>96.9</td>
</tr>
<tr>
<td><strong>Quality of Health Care in previous 12 months</strong></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>1.6</td>
</tr>
<tr>
<td>Fair</td>
<td>5.6</td>
</tr>
<tr>
<td>Good</td>
<td>32.2</td>
</tr>
<tr>
<td>Very Good</td>
<td>42.2</td>
</tr>
<tr>
<td>Excellent</td>
<td>18.3</td>
</tr>
</tbody>
</table>

Source: 2014 nationally-representative survey of US Adults. Author’s compilation

We observe a significant association between trust in confidentiality and whether patients had ever withheld information from their doctor (OR = 0.20, Table 2). However, no significant associations were found between trust in provider honesty and competence and having ever withheld information. In a model predicting patients’
expectation that it is very important that providers share their health information electronically, we find a significant association with trust in physician competence (OR = 2.87), but no significant relationship with trust in physician confidentiality or honesty. Finally, in a model predicting patients’ expectation that it is very important that they find out who has looked at their medical records, we find a significant association with trust in physician confidentiality (OR = 0.59), but no significant association with trust in physician honesty or competence. No model showed a significant association between overall trust and the three dependent variables. All models control for race, gender, education, and age.

Table 2. Odd-ratios for Logistic Regression of Trust in Physician dimensions on Patient Behavior and Expectations about Information Access and Sharing, Odds-ratios [95% CI]

<table>
<thead>
<tr>
<th></th>
<th>Ever Withheld Information because concerned about privacy</th>
<th>Very important that HC providers share my information electronically</th>
<th>Very important that I find out who has looked at my medical records</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Trust in physician</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.60</td>
<td>0.45</td>
<td>1.94 *</td>
</tr>
<tr>
<td></td>
<td>[0.32 – 1.1]</td>
<td>[0.15 - 1.4]</td>
<td>[1.2 – 32]</td>
</tr>
<tr>
<td>Honesty</td>
<td>2.45</td>
<td></td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td>[0.78 - 7.7]</td>
<td></td>
<td>[0.69 - 2.5]</td>
</tr>
<tr>
<td>Competence</td>
<td>3.53</td>
<td></td>
<td>2.87 *</td>
</tr>
<tr>
<td></td>
<td>[0.77 - 16.2]</td>
<td></td>
<td>[1.1 - 7.8]</td>
</tr>
<tr>
<td>Confidentiality</td>
<td>0.20 ***</td>
<td></td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>[0.10 - 0.42]</td>
<td></td>
<td>[0.48 - 2.2]</td>
</tr>
<tr>
<td>With Controls</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>F-test</td>
<td>3.9 ***</td>
<td>3.3 ***</td>
<td>4.6 ***</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>14,700</td>
<td>17,697</td>
<td>14,700</td>
</tr>
</tbody>
</table>

Source: 2014 nationally-representative survey of US Adults. Author’s compilation.

Discussion

Trust is an important determinant of patients’ behaviors and experiences of healthcare,9,10,12,13 and here we show how it is associated with information sharing. We provide evidence that specific dimensions of trust in physicians, namely, trust in confidentiality and competency of providers, influence patients’ behaviors and expectations for information sharing in health care. This is in contrast to much previous work which considers trust to be a singular, generalized determinant of patient behavior. We find that specific dimensions of patient trust in physicians have unique associations, namely, trust in confidentiality is associated with lower odds of having ever withheld information and reporting concern about who is accessing their health record information. In addition, patients with higher levels of trust in physician competency have greater odds of agreeing that it is important for providers to share health information with each other. Encouraging patients’ health information sharing, and reducing patients’ withholding, may increase the accuracy and quality of health data available in health IT systems.

This work demonstrates not only that trust in providers is essential for information sharing, but that different dimensions of trust matter differently. This finding further supports previous work showing trust is situational12. Patients’ trust is not a singular, generalized phenomenon, instead it is a nuanced series of relationships based on specific expectations and behaviors. Moreover, we show that the individual dimensions of trust are particularly important to study, as each contributes in different ways to the patient experience and the physician-patient relationship.
Additionally, patient information disclosure and concerns about access are also multifaceted and nuanced. Patients are concerned about who accesses health data and why the data are used\textsuperscript{20}. Perceptions of how data are used may also influence patients’ trust\textsuperscript{20}. Trust in providers is associated with the quality of the doctor-patient relationship. Subsequently, dimensions of trust may play different roles in ensuring good communication between patients and providers, in patients’ use of healthcare services, and for adhering to treatment\textsuperscript{21,22}.

Given systematic disparities in patient trust, in which racial and ethnic minorities have lower trust in physicians\textsuperscript{13,14}, while patients with devalued health characteristics have lower trust in physician confidentiality than others\textsuperscript{12}, further research is needed to examine how these factors may intersect with the dimensions of trust and their relationships to health information disclosure and concerns as shown here. This paper controls for patient age, gender, race, ethnicity, self-rated health, insurance status, and quality of health care but further study is needed to disentangle the relationships and possible interactions between and among them, dimensions of trust and health information sharing. Further, we have little understanding of how features of health IT applications interact with dimensions of patient trust. Health IT application design may unintentionally diminish trust or exacerbate rather than reduce privacy concerns that hinder use of beneficial health IT

These critical gaps in knowledge mean practitioners are missing when some patients, particularly the most vulnerable, may be more likely to have incomplete or inaccurate health data. The consequences of incomplete data include less effective treatment, more potential medical errors, and possible bias in treatment due to lack of data-representation in decision-making algorithms. Similar to a lack of representation in clinical studies, when patients’ data is absent, practitioners risk not learning which patients may most benefit from precision health treatments.

This study has a number of limitations that must be considered. First, data for this analysis come from a cross-sectional sample collected in 2014. Expansions in health IT use since then, as well as the evolution of available technologies, may change patients’ experiences and perceptions of technology. Additionally, qualitative research would allow more in-depth understanding of how different dimensions of trust influence patients’ behaviors and expectations about physicians, health IT, and privacy.

**Conclusion**

New information technologies in health care have the potential to enhance care quality and access, yet the widespread information sharing enabled by them also create new risks and concerns. Patients concerned about privacy may be less likely to disclose health information to providers and may be concerned about how their information is being used and accessed within the health care system. While trust is known to be an important factor in health care delivery and doctor patient relationships, previous work has not shown how specific dimensions of trust in physicians contribute to patient information sharing behavior or concerns. Here we show that one particular dimension of trust, trust in provider confidentiality, is associated with patients’ withholding information from clinicians and concerns about who is accessing their records, while a different dimension, trust in physician competency, is associated with agreeing that providers should share information. This is an important contribution given the increasing reliance on databases of patient information to tailor the delivery of care.

Ensuring that new technologies and their benefits are available to all patients, especially the most vulnerable, is necessary to prevent health disparities from widening as new effective health information technologies come in to widespread use. What is needed is an understanding of how dimensions of patient trust matter for health IT use and endorsement, whether particular dimensions of trust alleviate privacy concerns, and how features of health IT influence patient privacy concerns or usage.
References


Usability Evaluation of an Adaptive Information Recommendation System for Breast Cancer Patients

Maia Jacobs, PhD1, Janice Hopkins, APRN, MBA2, Matthew Mumber, MD2, Elizabeth Mynatt, PhD3
1Harvard University, Cambridge, MA; 2Harbin Clinic Cancer Center, Rome, GA; 3Georgia Institute of Technology, Atlanta, GA

Abstract

We report on the usability of a mobile application, MyPath, that connects patients with personalized information based on their diagnosis and care plan and adapts over time as they progress through the cancer trajectory. We conducted usability tests with cancer survivors and health professionals, measuring three usability factors which could be affected by adaptive content: learnability, errors, and effectiveness. Our results indicate that the adaptive information did not obstruct usability of the system. Participants identified several strengths of the application, including the integration of clinical and non-clinical information, the segmentation of a large information set to reduce mental burden, and the inclusion of multiple media types to accommodate different learning styles. Participants also identified potential barriers to use and offered ideas for future developments. We share how we integrated this feedback into the MyPath system design and reflect on lessons for future personal health information systems.

Introduction

Following a cancer diagnosis, access to informational resources can improve patients’ engagement in their care and overall quality of life1,2. Information about the disease can increase patients’ confidence in their ability to cope with the effects of the disease at home1,3. Most often, patients will rely on health professionals and printed materials for information, though use of online health information has increased over the years4,5. Notably, each of these information sources have limitations that affect their usability and accessibility. For example, patients can be hesitant to raise questions to health professionals6,7. Printed materials can be difficult to organize, particularly as the collection grows over time, therefore making the information difficult to recall or retrieve later in the cancer trajectory, when the information could be best put to use8. Finally, patients have varying degrees of trust in online health information as these resources can be inaccurate and unreliable9.

We are exploring ways to use ubiquitous, mobile technologies to connect breast cancer patients with trusted health information while overcoming these existing barriers to information access. Prior patient-facing cancer management interventions that focus on informational support have been associated with improved symptom management10, increased confidence and engagement in care11,12, and reduced symptom severity and distress10,13. These systems include various features to support patients, including symptom monitoring10,11,14, decision support15, and health management advice12, which can be tailored based on user input14.

While these systems have demonstrated a number of positive effects, they often support a subset of patients’ information needs. Many existing systems focus on a particular moment in the cancer trajectory, such as during chemotherapy or radiation therapy. However, patients’ information needs both persist and adapt over time, as they transition from diagnosis through survivorship16. In addition, current systems tend to focus on helping patients manage the physical effects of cancer. Research consistently finds that patients’ information needs extend beyond physical health, and that patients often feel they do not receive enough information to effectively cope with the disease1,17. Finally, while a small set of systems tailor content to the individual, we have not found examples of automatically adapting content based on an individual’s care plan. Such adaptations can provide patients with timely information to help them prepare for upcoming changes in care. Transitional periods can be vulnerable times in the cancer trajectory which alter patients’ self-management abilities18.

We are working with a team of oncologists, cancer navigators19, cancer survivors, and computer scientists to design a novel patient-facing health management application. This application, MyPath, provides patients with comprehensive health information, spanning beyond physical health information. The information adapts over time based on an individual’s care plan and treatment dates, as well as user input of current information needs.
We are employing a user-centered design process to develop a usable application that provides patients with information that aligns with their particular support needs and preferences. User-centered design involves a systematic process for including the intended users (e.g., patients and healthcare professionals) in the design and testing of interactive health technologies, to ensure that new tools are important and usable to the intended users.20 A core component of user-centered design, and one of the most influential components of the design process, is usability testing.21 In usability testing, a small number of potential users of a system interact with a prototype, with the goal of identifying usability problems before the application is built.22 While usability tests have been used in the design of tailored informational support systems,23 we did not find any prior studies that specifically test the effect of adaptive content on usability. We assess three usability factors: learnability, errors, and effectiveness. Prior studies have shown that adaptive content can influence each of these usability factors.24, 25 In addition to measuring usability factors, we used a think aloud protocol, a method often used within usability assessments, to understand users’ expectations and reactions towards the technology.21 Thus, the objectives of this study were to measure the effectiveness, learnability, and error rates of adaptive information recommendations, identify usability issues through participant feedback, and inform the design of future systems that tailor content over time.

Method

MyPath Description
MyPath is a mobile application that connects a patient with a trusted and comprehensive set of information related to cancer management, which adapts over time as the patient progresses through the cancer trajectory. This design goal was supported by our empirical work with cancer navigators, understanding the different types of support and information needs across the patient population, as well as our work with breast cancer survivors creating journey narratives to assess unmet support needs.26 This work highlighted the broad and changing needs of breast cancer patients as they progress through multiple phases of care, and the need for support systems that reach across multiple phases of care, helping patients to cope with current challenges and prepare for future changes to their care.

MyPath’s database includes over 300 information resources from American Cancer Society, breastcancer.org, and cancer.net. We included these three sources based on oncologists’ recommendations for trusted online health information. MyPath’s information recommendations adapt in two ways. First, there is a user survey that may be completed at any time to immediately receive information related to current questions or challenges. The survey asks the user to select which challenges they are currently dealing with, and includes 58 possible responses. Survey items are based on the NCCN distress thermometer and amended with additional challenges captured in our research with cancer survivors.26 Metadata within the MyPath database associates specific survey responses with relevant informational resources. Thus, when a patient selects a particular response in the survey, such as ‘Nausea’, resources related to that subject will immediately appear in the application when the survey response is submitted. While the research team created the metadata tags, multiple oncologists and oncology nurses validated that resources were appropriately characterized and tagged by the application.

In addition to tailored content based on survey responses, MyPath automatically recommends information as patients prepare for and encounter new treatments. Each MyPath user has a profile that includes all treatments and treatment dates. These dates are used to signal a transitional event that generates recommendations based on the patients’ treatment path. Metadata in the database connects transitional events, (e.g. surgery, radiation therapy, chemotherapy, hormonal therapy) to relevant resources. Resources appear one week before a new treatment, or at the end of a treatment, to help a patient learn what to expect and prepare for this change. For example, if a patient is receiving chemotherapy, then the system will identify the start date minus seven days as a transitional event, and on this date the user will see new resources associated with chemotherapy, which help them learn about the treatment, what to expect, and how to prepare.

MyPath Prototype
Once we established MyPath’s features, we then worked on multiple design options using low-fidelity prototypes to generate a proof of concept and elicit initial feedback from members of the research team. This iterative process led to the creation of a high-fidelity prototype. The prototype was created using the Justinmind prototyping software and displayed on Verizon Ellipsis 8 tablets. The prototype includes four sections: the home page, the information recommendations, the resource view, and the user survey (Figure 1). Within the prototype, participants were able to
interact with all buttons and hyperlinks, allowing them to freely explore the system. When selecting one of the PDF resources, the prototype would link to the original online source, though these resources will be available offline in the final application.

Participants
We enrolled breast cancer survivors who had recently completed treatment. Access to a high number of cancer survivors was difficult as the cancer clinic serves a large, rural geographic region, with many survivors living a far distance from the clinic. Therefore, we also included cancer navigators and oncology nurses, who are highly knowledgeable of common questions and informational needs of patients throughout the cancer trajectory.

Usability task development and procedures
Due to the limited work evaluating adaptive interfaces, particularly in a health context, the usability assessments for MyPath were critical for understanding any complications caused by adapting the content over time. We developed three tasks in the usability test to demonstrate changes to the content over time, summarized in table 1. In the first task, participants were presented with information that would be displayed at the time of diagnosis. Before interacting with the prototype, participants were asked to reflect on either their own information needs during this time, or a common information need of their patients. Participants were then instructed to open the application and find two specific resources relevant to the issue they identified that would be helpful to a newly diagnosed patient. To accomplish this task, participants needed to 1) open the application, 2) select a category, 3) select one of the information pages to read, and 4) repeat steps 2-3 to find an additional resource. In the second task, we presented participants with health information relevant to a patient who was beginning radiation in one week. Again, we asked participants to first reflect on their own information needs during this time, or the needs of their patients. Participants were then asked to find two resources that address those information needs, thus repeating the same actions described in the first task. Notably, between the first and second tasks the facilitator updated the prototype to a second homepage view which had new resources added, thus visually mimicking the adaptive feature of the MyPath system. Participants were not told that the information recommendations would be updated.

In the final task, we asked participants to imagine they were in the middle of treatment and to complete the user survey. While the final MyPath application will suggest resources based on these selections, the prototype did not include this feature. Therefore, we asked participants to describe the types of information they would expect to see appear after completing the survey.
Table 1: Description of tasks included usability tests.

<table>
<thead>
<tr>
<th>Task No.</th>
<th>Task Description</th>
</tr>
</thead>
</table>
| Task 1   | (A) Think about a question you (or one of your patients) had following diagnosis. Try to find a resource within MyPath that addresses this question.  
(B) Try to find a second resource that addresses this question (from Task 1). |
| Task 2   | (A) Think about a question you (or one of your patients) had a week before beginning radiation therapy. Try to find a resource within MyPath that addresses this question.  
(B) Try to find a second resource that addresses this question (from Task 3). |
| Task 3   | Think about a question you (or one of your patients) had during radiation therapy. Complete the MyPath survey and identify one resource that addresses this question. |

Data collection and analysis

All usability sessions took place at the Harbin cancer clinic, in a private meeting room. The study involved a one-hour, one-on-one session with a member of the research team, who has a background in human-computer interaction and over seven years of experience running usability studies. While the facilitator was a member of the team who designed the application, participants were not made aware of this, in order to reduce possible influences on participants’ responses. Each session included the usability test, a post-test questionnaire, and interview. To ensure consistency, the facilitator used a script throughout the study to introduce participants to the study goals and procedures. The usability study was approved by the Georgia Institute of Technology’s Research Ethics Board.

We captured participants’ task performance, usability scores, and post-study feedback in order to assess three usability heuristics: learnability, errors, and effectiveness. Learnability refers to the increase in users’ ability to efficiently accomplish tasks over time, as they become familiar with a new system. Prior research has shown that adaptive interfaces can be unpredictable, thus affecting a system’s learnability. We expected that if the adaptive recommendations in MyPath impaired the system’s learnability, we would not see a reduction in task performance time between the first and second tasks. We measured task performance time as the elapsed time between the participant opening the application and the participant identifying (audibly) two useful resources.

Error rates are another metric used to identify significant usability issues in a design. Errors, noted by the facilitator, included an event in which a participant navigated to a page that did not match their intention (for example, looking for treatment information in the Social Support page), or participants were unable to locate an intended screen (such as the user survey). We used a think aloud protocol to capture participants’ intentions while using the system. In a think aloud protocol, participants are encouraged to think out loud as they interact with the application, so that the researcher may note when participants feel confused or experience difficulties. This method is useful for capturing obstacles that participants face throughout the usability assessment.

Similar to prior research, we included measures of effectiveness in the usability test to ensure that the application features were considered useful by participants. We both objectively and subjectively measured the system’s effectiveness. We used task completion rates as a measure of effectiveness, as a task was only considered complete if participants located resources that they considered useful for coping with the information challenges described prior to the task start. Following the usability test, we also asked users to rate on a 5-point Likert scale if they felt the resources included in MyPath would be useful for addressing their information needs (1=not at all useful, 5=very useful).

Finally, we captured post-study feedback using semi-structured interviews. All interviews were audio recorded. Two researchers transcribed the recordings and conducted a thematic analysis of the qualitative data. This analysis included first open-coding the transcripts independently, and clustering segments to develop theme concepts. The two researchers then compared codes, discussed discrepancies, and revised the coding scheme. One researcher then did a final review of all transcripts, verifying themes across the full dataset. The research team then met to discuss the final themes and identify necessary design changes.
Results

Participant characteristics
We enrolled 11 participants in the usability evaluation, including four breast cancer survivors and seven healthcare professionals. The healthcare professionals had an average of 16 years of experience working with cancer patients, ranging from 1 to 45 years. All participants were female, with an average age of 52 years. Participants’ average technology comfort score was 3.8 on a 5-point Likert scale (1 being very uncomfortable, 5 being very comfortable).

Learnability
We compared the completion times of the first and second usability tasks to assess learnability. We use the completion times of ten participants, as one participant was unable to complete the second task (see Effectiveness discussion below). The average completion time of the second task (mean=33.1, SD=10.5) was significantly lower compared to the first task (mean=58.9, SD=9.8) (t=5.0, df=9, p < .001).

Errors
Participants performed a total of 33 tasks in the usability tests, summarized in table 1. Across these tasks, we identified 13 errors. The majority of errors (70%) occurred in the third task. The most common error was incorrectly navigating to the user survey, with five participants unable to locate the survey button during the assessment. Three participants also had difficulty selecting the checkboxes within the survey, due to the widgets being too small.

Effectiveness
The task completion rate was 97%, with one participant not completing the second task due to not finding information that addressed her described information need. In this case, the participant was looking for transportation support offered within the community. The lack of local resources also came up frequently in the participants’ feedback, as we discuss below. All other participants were able to find multiple resources that they felt would be useful for addressing their described information needs.

During the first two tasks, participants were each asked to identify four informational resources to address patient questions or challenges. Participants were free to view any categories that they believed would include information relevant to their described information need. During these two tasks, participants identified 43 resources which spanned all of the MyPath categories. The most commonly viewed category was Treatments (n=18), followed by Day to Day Matters (n=8), Emotional Support (n=7), Health and Wellbeing (n=6), Overview (n=2), and Social Support (n=2). The range of information needs described by participants, and categories viewed, validates the importance of comprehensive informational support.

The average usefulness score from the post-test questionnaire was 4.2 out of 5. No participant rated the system’s usefulness below a 4, indicating an agreement in the expected utility of the application. However, in the post-test interviews, described below, participants highlight two issues that could impede on the system’s overall effectiveness.

Participant Feedback
We identified six themes during the analysis of the interview data. These themes fell into three categories: system strengths, anticipated barriers, and design recommendations. Table 2 shows a summary of these themes and the number of participants who mentioned each theme at least once during the interview.

All participants believed MyPath would be helpful for future patients. Most often, participants said that they liked that MyPath included a breadth of information, addressing needs spanning beyond physical symptoms and side effects. One participant shared that her priorities did not always align with the information that doctors shared, and she was able to find useful information in the prototype:

I think sometimes physicians, because they do it every day, they don’t realize that there’s a lot of things that are not discussed with patients. Day to day matters, social support. [Pointing to MyPath] I would have looked through here for how do I support my family. I’m a mom. I’m a working mom. What’s there for my family? - Breast cancer survivor

Seven participants commented that MyPath recommended short lists of information. Participants liked that the application would populate with small, digestible sets of information during treatment transitions. Both cancer survivors and health professionals noted that large information sets often become overwhelming and ignored:
Table 2: Participant feedback themes, and the number of participants who mentioned each theme at least once during the post-usability study interview.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Total Mentions</th>
<th>% of survivors</th>
<th>% of health professionals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strengths</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to non-clinical resources</td>
<td>8</td>
<td>75%</td>
<td>71%</td>
</tr>
<tr>
<td>Short lists</td>
<td>7</td>
<td>75%</td>
<td>57%</td>
</tr>
<tr>
<td>Multiple media types</td>
<td>6</td>
<td>75%</td>
<td>43%</td>
</tr>
<tr>
<td><strong>Barriers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Navigation challenges</td>
<td>5</td>
<td>25%</td>
<td>57%</td>
</tr>
<tr>
<td>Different attitudes towards information seeking</td>
<td>4</td>
<td>50%</td>
<td>28%</td>
</tr>
<tr>
<td><strong>Design Recommendations by Participants</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integration with in-person support services</td>
<td>7</td>
<td>50%</td>
<td>71%</td>
</tr>
</tbody>
</table>

*I like that the lists are short. Sometimes if we give them too many resources they don’t even want to read them.* - Oncology nurse

*I was bombarded by brochures [when I was diagnosed]. This seems like these things are sent directly, specifically to me by the creators.* - Breast cancer survivor

Finally, six participants said that they liked that MyPath included multiple media types. While most of the information was in html or pdf format, the database also included videos and discussion forums. Both the health professionals and cancer survivors thought this would be useful, due to patients’ different learning styles:

*I see there are several videos, and because people learn in different ways, that’s something that I think might be pretty good.* - Breast cancer survivor

[Pointing to a video] *I would rather listen than read.* - Breast cancer survivor

Participants also noted barriers that they believed hurt the usability of the system. Five participants commented on navigation challenges they encountered, or would expect to encounter, which further substantiated the errors we identified in the usability test. Participants were particularly concerned about how to navigate to new resource recommendations:

*Could it highlight the new resources, and a message that said please tap the highlighted field? Could it direct me in some way to go there?* - Breast cancer survivor

While participants were overall positive about the potential usefulness of the MyPath application, four participants also noted that patients respond differently to health information. Participants cautioned that this type of application may not be useful for all cancer patients. As one cancer survivor shared, some people choose not to seek out information:

*There is so much information out there. Some of it is not good. Some people want to know everything. I am not that kind of person.* - Breast cancer survivor

Finally, participants shared ideas for future information systems. The most common recommendation was to incorporate information about patients’ care team and local resources within the patient’s community. Both health professionals and cancer survivors saw this application as an opportunity to help patients connect with in-person support systems:

*You could do a little sentence at the end. If your distress level is 4 or higher, please share this information with your healthcare team.* - Cancer navigator

*I would like more about what is there actually for ME in this community? Is there transportation help? Is there financial help?* - Breast cancer survivor
Discussion

We are developing MyPath, a mobile health application that connects cancer patients to tailored health information based on user input and their care plans, to assess opportunities and barriers of using adaptive systems to increase cancer patients’ access to health information. Thus far, little is known about how dynamic content within an adaptive intervention may affect the system’s overall usability. In this study, we assessed three usability factors that we hypothesized could be influenced by dynamic content: learnability, errors, and effectiveness. Our results indicate that adaptive health information presents a useful approach for making large information sets more manageable for patients, while also raising important usability issues requiring design alterations.

Our first goal within this usability study was to determine if the automatic addition of content over time would interfere participants’ ability to learn how to use the application and find relevant content. Usability studies within other contexts have suggested that adaptive interfaces can interfere with one’s ability to learn a system’s features. To test the system’s learnability, we compared the task completion times of the first and second tasks in the usability tests, which require participants to complete the same set of interactions, but with new information recommendations added to application prior to the second task. We found that task completion times were significantly lower in the second task, suggesting that dynamic content did not interfere with participants’ ability to find relevant information.

We found that the majority of errors were related to navigating to the user survey. Further, participants were concerned about navigating from the user survey to new resources that were added as a result of the survey submission. These errors highlight the importance of adding navigation support to help participants move between system features. Prior work exploring the design space for adaptive systems have discussed the importance of helping users to navigate these complex interfaces, especially if individuals interact with the system infrequently. Our study shows that when recommending health information to patients, adaptive systems need to visualize when, why, and where new information recommendations have been added.

Participants’ high task completion rates and usefulness scores suggest a consensus that participants believe MyPath will be helpful for breast cancer patients. However, during the interviews, participants did suggest that integrating information about one’s care team and other local resources would be important for patients and increase the system’s utility. In-person resources, such as support groups, self-management programs, and peer navigation services, can improve a patient’s quality of life throughout the cancer trajectory. However, a lack of information about available resources can limit patients’ ability or willingness to access local services. Participants in this study frequently commented that adaptive information systems could be help bridge these gaps, connecting patients with in-person support services.

Based on this feedback, we made a number of changes to MyPath. First, we added a local resources category that includes links to the cancer clinic and other support services for cancer patients. Second, we moved the user survey to a more prominent location on the homepage and added a text label. Third, we added a number of navigational support features. We changed the labels within each category to more accurately represent why new resource recommendations were added (labels originally only included the date the resources were recommended). We also included animation to show which categories include newly recommended resources, as well as messages with a description of new resource recommendations that are added after a survey submission or at the start of a treatment transition. Fourth, we added a question to the user survey that asks participants how much distress they are experiencing. When a patient selects a distress level above 4, a message appears suggesting they get in touch with a healthcare professional. These changes are depicted in Figure 2.

The results from this study point to design guidelines for adaptive, patient-facing tools. As noted above, adaptive tools should include clear navigational support for users and may be useful in connecting patients with in-person support when necessary. Participants in this study also raised the issue of cancer patients having different attitudes towards information seeking. We see this issue as an important opportunity for future research. Many cancer patients use information avoidance as a coping strategy. Open questions remain about how health management tools may adapt over time to accommodate this behavior. Future research addressing this question could support the development of informational support that is beneficial for a broader patient population.
Limitations

Our usability assessment included a number of limitations. We worked with a small number of participants, and a larger participant group may identify more usability issues, though research suggests that this participant size typically captures a majority of the usability issues. We chose not to include patients who were recently diagnosed or in the middle of treatment. Often these patients are traveling far distances to come to the cancer clinic for treatment, and we wanted to be respectful of their time and mental well-being. We opted to recruit healthcare professionals and cancer survivors who had progressed through diagnosis and treatment, who would be able to reflect on how patients’ information needs change over time. However, it is possible that the emotional burden of the diagnosis and treatments will influence the effectiveness of the tool. We will use a deployment study to assess real-world barriers to using adaptive health information recommendations throughout the cancer trajectory.

Conclusion

Adaptive content can have unforeseen consequences on the usability of an application. The findings from this study emphasize the importance of incorporating navigation support within adaptive systems. Participant feedback underscored the importance of visible system recommendations, with clear explanations of when and why new recommendations are being presented. Despite navigation challenges, participants consistently stressed the potential benefits of adaptive health information for breast cancer patients, such as helping to make large information sets more manageable and less overwhelming, while also revealing an important opportunity to use such tools to help patients connect with healthcare professionals and local support services. These findings may useful in designing technologies for other illness trajectories in which patients’ information needs change over time, including other types of cancer and illnesses with intermittent serious episodes, in which patients’ and caregivers’ information needs may significantly change over time. We also expect that as adaptive patient-facing tools become more common, this approach to usability testing may be useful for assessing how dynamic content influences usability in other contexts.

Acknowledgments

We would like to thank our participants for volunteering their time to provide their crucial feedback. We would also like to thank the Cancer Navigators of Rome, Lena Crooker, and all of our partners who are a part of the Rome, GA healthcare community. This work is supported by the National Institute of Health (RO1 CA195653).
References


4. Fox S, Purcell K. Chronic Disease and the Internet; 2010.


Machine Learned Mapping of Local EHR Flowsheet Data to Standard Information Models using Topic Model Filtering

Steven G. Johnson, PhD¹, Lisiane Pruinelli, PhD, MS, RN¹,², Bonnie L. Westra, PhD, RN, FAAN, FACMI¹,²

¹Institute for Health Informatics, University of Minnesota, Minneapolis, MN;
²School of Nursing, University of Minnesota, Minneapolis, MN

Abstract

Electronic health record (EHR) data must be mapped to standard information models for interoperability and to support research across organizations. New information models are being developed and validated for data important to nursing, but a significant problem remains for how to correctly map the information models to an organization’s specific flowsheet data implementation. This paper describes an approach for automating the mapping process by using stacked machine learning models. A first model uses a topic model keyword filter to identify the most likely flowsheet rows that map to a concept. A second model is a support vector machine (SVM) that is trained to be a more accurate classifier for each concept. The stacked combination results in a classifier that is good at mapping flowsheets to information models with an overall f2 score of 0.74. This approach is generalizable to mapping other data types that have short text descriptions.

Introduction

Healthcare data is now readily available in electronic form thanks to the continued adoption of electronic health records (EHR). Never before have we had easy access to so much data that has the promise to improve patient outcomes. But the secondary use of this data for analysis and sharing across organizations is stymied because we don’t have robust standard and formal information models to support analyzing and comparing the data¹. An information model is a formal structure for representing the clinical information in the EHR and includes data elements, relationships between the elements, and rules that the data elements should satisfy². While efforts such as the Observational Medical Outcomes Partnership (OMOP), Patient-Centered Outcomes Research Institute (PCORI), Fast Healthcare Interoperability Resources (FHIR) and Clinical Information Modeling Initiative (CIMI) have made good progress toward standardized information models³–⁶, those efforts have not given attention to data captured in the delivery of nursing and other inter-professional areas.

Much of this data is semi-structured and captured as “flowsheet” data⁷. Flowsheet data is used by many healthcare organizations to record custom and non-standardized information into the EHR. It is arranged as a spreadsheet like data entry with rows representing the different data types and columns representing time periods to record observations of the rows. There is usually not an information model that the EHR builders use to model flowsheet data. Flowsheet data is particularly difficult to model because most of it is not coded to standard terminologies like Logical Observation Identifiers Names and Codes (LOINC) or Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT). Therefore, flowsheet data can’t be mapped using tools such as Regenstrief LOINC Mapping Assistant (RELMA)⁸ or MetaMap⁹. Flowsheet rows (data types) typically have very little information that describe the data that is being captured.

Recent work has resulted in the development of information models in 10 key areas important to nursing¹⁰. Eight organizations formed an Information Model Validation Work Group (IMVWG) to evaluate and validate these models across their organizations. Two of the models are nearing completion for validation across organizations. One of the models, the Pain Information Model (Pain IM), has been validated and published¹¹. However, a significant problem still remains even after standard information models have been made available for use. The problem is that each organization then has the unenviable task of mapping the standard information models to the data in their local EHR system. This can be a daunting, time-consuming and error prone task, particularly since organizations may have multiple flowsheet rows for the same type of data element. Often these are created for different units, disciplines, or types of settings, and change over time with upgrades to the system.

In the EHR systems of the eight organizations in the IMVWG, flowsheet data types are described using two 90-character fields. The first field is the flowsheet data type’s internal name (called a flo_meas_name), and the second is
the text that is displayed on the data entry screen (called a flo_disp_name). Each flowsheet data type also has a unique identifier, which can be different in each organizations’ EHR (called a flo_meas_id), but those identifiers are not guaranteed to be unique across different systems. For example, different systems can use flo_meas_id=12345. The first may use it to represent “Blood Pressure” and the second can use it to represent “Heart Rate”. Furthermore, the descriptions for the same concept (i.e. “Pain Rating 0-10 Scale”) may use the text “Pain Scale” at one health system and “Pain Rating” at another.

Once a standard information model has been developed (for example, the Pain IM), each of the organizations in the IMVWG assigned staff to map their flowsheet data types to the information model concepts. This was a time-consuming manual process that needed to be done after every model was developed. It required staff to search through flowsheet row descriptions and find the rows that they believed mapped to the concept in the information model. The problem was worse for new organizations that wanted to join the IMVWG. To get the benefit of the work group, they would have to manually map their flowsheet data to all of the information models, which required a large initial commitment of time and resources on their part. In order to make that initial mapping task less daunting, the IMVWG sought a way to automate the mapping process. The work group looked at natural language processing (NLP) and machine learning as possible approaches. Approaches using traditional NLP techniques were deemed unlikely to work because the flowsheet row descriptions are very short, are not sentences and don’t have traditional parts of speech. There has been work on mapping short descriptions as part of the caBIG project, but those tools were focused on rank ordering matches and supporting researchers in manually mapping between common data elements.

The purpose of this study was to develop and implement a technique for mapping local EHR flowsheet data types that have short descriptions to standard information models using a machine learning approach. The technique will be illustrated using the Pain IM.

**Methods**

The overall approach has four steps:

1. Develop a model using the IMVWG manually mapped flowsheet rows as training data for a machine learning algorithm
2. Evaluate the model performance using the f2 score
3. Validate the model predictions by having researchers review the results
4. Update the incorrect mappings from the IMVWG and then re-train and re-evaluate the model

Machine learning is a powerful technique for developing models. The basic approach for mapping the Pain IM to local EHR data was to obtain examples of correctly mapped flowsheet rows to concepts in the Pain IM and use that data as training data to build a machine learned model that can label new flowsheet rows to map it to the right concept in the Pain IM. A portion of the Pain IM is shown in Figure 1.
One task of the IMVWG was for each organization to manually map their organization’s flowsheet rows to the Pain IM. As a result of that work, eight organizations mapped 1,837 flowsheet rows to the 103 concepts in the Pain IM. These previous manually developed mappings served as training data for the machine learning models.

**Model Development**

The method selected for development of a machine learned model was influenced by the flowsheet data structure and volume. In the EHR used by most of the workgroup participants, flowsheet rows are described using 60-120 characters of text. Because the EHR builders had to fit quite a bit of information into a small space, much of the description makes use of abbreviations and short phrases instead of full sentences to describe the flowsheet data types, so the words in the descriptions act like keywords.

The initial attempt to create a machine learned model for predicting labels used a term frequency-inverse document frequency (tf-idf) approach. Each short description (the 60-120 characters of text) that was defined in the EHR that described each flowsheet row is considered a document. Tf-idf creates a bag-of-words count of each of the terms that occur in these descriptions, but gives a higher weighting to rare terms and a lower weight to common terms. This results in word vectors that are the set of features covering all of the flowsheet row descriptions.

For the training, depending on which organization’s data was left out for testing, there were typically 100,000 negative examples (unmapped flowsheet rows) and approximately 500 positive examples (flowsheet rows that were mapped to concepts in the Pain IM), which is an extremely unbalanced dataset for machine learning. And in addition to just looking at each individual word for the bag-of-words parsing, tf-idf was configured to also included all of the unique two- and three-word phrases (bigrams and trigrams). This led to having models with 800,000 to 1,000,000 features.
Training a support vector machine (SVM) using that many features on the unbalanced dataset would be very slow and would not result in good model performance.

There are a number of approaches in machine learning that attempt to address the imbalance including random under- or over-sampling\textsuperscript{14} and SMOTE\textsuperscript{15} that creates synthetic samples. These approaches have disadvantages in that they either add noise or remove potentially useful information. Some studies have found that using model ensembles can intelligently identify sub-samples that will improve model performance when using imbalanced data\textsuperscript{16}. Therefore, a stacked model approach\textsuperscript{17} was pursued where the first model would filter the flowsheet rows to identify rows that should definitely not be categorized as a Pain concept and the second model would be a SVM to more accurately classify a flowsheet row as a Pain concept or not. The first stage of the stacked model significantly reduced the negatives using a topic modeling filter based on the TextRank algorithm\textsuperscript{18}. A topic model was developed for each concept in the Pain IM. The top $N$ topics were selected for that concept (enough to ensure that 100\% of the positive training data was included in the topic filter). This topic filter was then applied to the training data and the second model, a SVM, was trained on the remaining data.

A separate topic model and SVM was built to classify each of the concepts in the Pain Information Model. The Pain IM has 103 concepts, so in the end there were 103 2-stage models run on the unmapped flowsheet data of a new organization to label and map their flowsheet rows. The SVM produces a score which is a probability that a label is correct for a particular flowsheet row. Each model (topic filter and SVM) was applied to the flowsheet rows, and the label from the SVM that produced the highest score (probability) was used to label the flowsheet row. The process is shown in Figure 2. The software was written in Python and used the gensim\textsuperscript{19} library for tf-idf and TextRank and scikit-learn\textsuperscript{20} libraries for the SVM models.
Figure 2. Process for mapping flowsheet rows using a stacked classifier for each Information Model concept

Model Evaluation

The performance of the model was evaluated using the f2 score. The f2 score is a combination of the recall and precision of the model, but it is biased toward recall. The formula for f2 is shown in Figure 3.

\[
F_2 = 5 \cdot \frac{\text{precision} \cdot \text{recall}}{4 \cdot \text{precision} + \text{recall}}
\]

Figure 3. The f2 score defined in terms of precision and recall

In our use case, we don’t mind having more False Positives (when the model labels a flowsheet row, even though it may not correctly map to a concept in the information model) so the f2 score is a good choice for model evaluation. The f2 score weights recall as twice as important as precision. Using a hyperparameter search, an SVM probability score threshold was found that maximized the f2 score across the entire set of SVMs.
In order to simulate the mapping process that occurs when a new organization joins the IMVWG, an approach was taken to train the model on 7 organizations’ mappings and leave one organization out to be used as the testing data to evaluate the model performance. This was done 8 times, leaving a different organization’s data out as the testing data each time.

**Model Validation**

The manually developed mappings from the workgroup were used as the “gold standard” to train the model. These mappings were known to have mistakes, such as a single flowsheet row mapping to the wrong concept or two different concepts in the IM. Or a flowsheet row might not be mapped to the any concept even though the concept existed in the Pain IM. So we also carried out a mapping validation process where the predictions of the model were manually reviewed by at least two researchers to determine if the original “gold standard” mappings were correct or if the prediction from the model was correct.

It would have been difficult to have the researchers manually review all 126,957 flowsheet row mappings from the eight IMVWG organizations. Instead, the researchers reviewed only the model predictions that differed from the IMVWG mappings. The differences fell into four Mismatch Categories:

1. A flowsheet row wasn’t mapped by the IMVWG, but the model predicted a new label that the researchers verified was correct.
2. A flowsheet row was mapped by the IMVWG (and verified correct by the researchers) but the model did not predict any label.
3. A flowsheet row was mapped by the IMVWG, but the model predicted a new label that the researchers verified was correct and better than the IMVWG.
4. A flowsheet row was mapped by the IMVWG, but the model predicted a label that the researchers verified was the wrong label.

Based on this review, the labels of the original IMVWG mappings were updated and the model training and testing process was performed again to determine the final model performance.
Results

The flowsheet row data from eight organizations was aggregated into a database containing information about 126,957 flowsheet rows in total. The model training process was run on 7 organizations’ data and the model was evaluated using the flowsheet data from the organization that was left out. The results of this process when using the original mappings are shown in Table 1.

<table>
<thead>
<tr>
<th>Org</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Rec</th>
<th>Prec</th>
<th>f2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>119</td>
<td>36</td>
<td>12,504</td>
<td>0.30</td>
<td>0.58</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>142</td>
<td>17</td>
<td>9,586</td>
<td>0.08</td>
<td>0.41</td>
<td>0.09</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>90</td>
<td>17</td>
<td>13,040</td>
<td>0.44</td>
<td>0.81</td>
<td>0.49</td>
</tr>
<tr>
<td>4</td>
<td>154</td>
<td>156</td>
<td>44</td>
<td>21,172</td>
<td>0.50</td>
<td>0.78</td>
<td>0.54</td>
</tr>
<tr>
<td>5</td>
<td>245</td>
<td>202</td>
<td>227</td>
<td>33,439</td>
<td>0.55</td>
<td>0.52</td>
<td>0.54</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>168</td>
<td>19</td>
<td>14,219</td>
<td>0.16</td>
<td>0.62</td>
<td>0.18</td>
</tr>
<tr>
<td>7</td>
<td>61</td>
<td>103</td>
<td>127</td>
<td>7,049</td>
<td>0.37</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>8</td>
<td>133</td>
<td>198</td>
<td>25</td>
<td>13,500</td>
<td>0.40</td>
<td>0.84</td>
<td>0.45</td>
</tr>
<tr>
<td>All</td>
<td>758</td>
<td>1,178</td>
<td>512</td>
<td>124,509</td>
<td>0.39</td>
<td>0.60</td>
<td>0.42</td>
</tr>
</tbody>
</table>

The f2 score ranged from 0.09 to 0.54 with an overall f2 score of 0.42. There were 1,690 FPs and FNs where the model and the IMVWG mappings did not match. Three researchers (two were clinicians) examined those mappings to determine the correct labels for the flowsheet rows. The Cohen’s Kappa was 0.80, which indicates good agreement between the researchers for what the correct labels should be.

Examples of IMVWG flowsheet labels and predicted labels are shown in Table 2. The table includes examples that were correctly mapped (TP) as well as examples of mappings from each of the four Mismatch Categories.

Table 2. Examples of Model Mappings and Mismatches

<table>
<thead>
<tr>
<th>IMVWG Label</th>
<th>Model Label</th>
<th>Flowsheet Row Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain Rating 0-10 Scale</td>
<td>Pain Rating 0-10 Scale</td>
<td>Pain Intensity (0-10)</td>
</tr>
<tr>
<td>Current Pain</td>
<td>Current Pain</td>
<td>Does the Patient Have Pain?</td>
</tr>
<tr>
<td>&lt;no label&gt;</td>
<td>Pain Rating 0-10 Scale</td>
<td>Pain Level</td>
</tr>
<tr>
<td>&lt;no label&gt;</td>
<td>Current Pain</td>
<td>Is child experiencing pain now?</td>
</tr>
<tr>
<td>Mismatch Category 2: Model did not predict any label</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current Pain</td>
<td>&lt;no label&gt;</td>
<td>Have you been feeling aches and pains?</td>
</tr>
<tr>
<td>Pain Rating 0-10 Scale</td>
<td>&lt;no label&gt;</td>
<td>Postop Pain Assessment</td>
</tr>
<tr>
<td>Mismatch Category 3: Model label is better than IMVWG label</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain Assessment Tool Used</td>
<td>Pain Rating 0-10 Scale</td>
<td>Pain Rating</td>
</tr>
<tr>
<td>Current Pain</td>
<td>Pain Assessment Tool Used</td>
<td>Pain Scale Type</td>
</tr>
<tr>
<td>Mismatch Category 4: Model did not predict correct label</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current Pain</td>
<td>Pain Rating 0-10 Scale</td>
<td>Presence of Current Pain</td>
</tr>
<tr>
<td>Pain Rating 0-10 Scale</td>
<td>Preferred Pain Assessment Tool</td>
<td>OT Seating and Wheelchair Mobility Pain Scale II</td>
</tr>
</tbody>
</table>
Using the updated labels, the models were retrained in the same manner by using 7 organizations data for training and then testing on the organization that was left out. This was repeated so that each organization was used as the test data once. The resulting model statistics are shown in Table 3. The f2 scores improved and ranged from 0.59 to 0.86, with an overall f2 score of 0.74.

Table 3. Model performance scores (f2) for each organization (updated labels)

<table>
<thead>
<tr>
<th>Org</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Rec</th>
<th>Prec</th>
<th>f2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96</td>
<td>38</td>
<td>48</td>
<td>12,527</td>
<td>0.72</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>2</td>
<td>111</td>
<td>35</td>
<td>18</td>
<td>9,593</td>
<td>0.76</td>
<td>0.86</td>
<td>0.78</td>
</tr>
<tr>
<td>3</td>
<td>114</td>
<td>43</td>
<td>14</td>
<td>13,048</td>
<td>0.73</td>
<td>0.89</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>245</td>
<td>69</td>
<td>14</td>
<td>21,198</td>
<td>0.78</td>
<td>0.95</td>
<td>0.81</td>
</tr>
<tr>
<td>5</td>
<td>408</td>
<td>169</td>
<td>162</td>
<td>33,374</td>
<td>0.71</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>6</td>
<td>104</td>
<td>81</td>
<td>37</td>
<td>14,215</td>
<td>0.56</td>
<td>0.74</td>
<td>0.59</td>
</tr>
<tr>
<td>7</td>
<td>112</td>
<td>38</td>
<td>127</td>
<td>7,063</td>
<td>0.75</td>
<td>0.47</td>
<td>0.67</td>
</tr>
<tr>
<td>8</td>
<td>277</td>
<td>53</td>
<td>9</td>
<td>13,517</td>
<td>0.84</td>
<td>0.97</td>
<td>0.86</td>
</tr>
<tr>
<td>All</td>
<td>1,467</td>
<td>526</td>
<td>429</td>
<td>124,535</td>
<td>0.74</td>
<td>0.77</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Discussion

The final model performance was quite good with an overall f2 score of 0.74. This approach will be very useful to new organizations that are added to the IMVWG. They will be able to get a high percentage of their flowsheet data types mapped automatically to the information models. The stacked models did a good job of finding the true positives (flowsheet rows that match the organization’s manual mappings) and true negatives (flowsheet rows that should not be labeled) and the false positives and false negatives were minimized. There are a number of findings from this research discussed below.

*Manual mapping is error-prone.* The first finding is that having human reviewers manually mapping flowsheet rows to concepts in an IM is error prone. There are many examples of mappings that were missed by organizations. When the authors reviewed the results, we found 388 additional flowsheet rows that could be mapped to a concept. This is not surprising since the clinicians in the organization who are performing the mapping can’t possibly be aware of all the places the concept is in their flowsheet data. They were given a tool that can do sophisticated Boolean search expressions using keywords, but that still relies on the skill of the mapper to remember all the ways a concept is represented in their flowsheet data and which keywords to use for a search. A better approach is to have the computer suggest mappings and then the human reviewer only needs to decide if it is a good mapping or not. The automated mapping techniques described in this study makes that approach possible.

*Automated mapping needs good training data.* The models did not perform as well when there were insufficient flowsheet instances present in training data. For example, only two of the organizations had a single flowsheet row named “Pain Level” mapped to the Pain Rating 0-10 Scale concept. When one of the organizations was used as the testing data, there was only one instance of “Pain Level” in the training data, which was not enough for the SVM to be able to consistently classify it to the concept of Pain Rating 0-10 Scale. On the other hand, there were many organizations that used phrases like “Pain Rating” and “Pain Scale”, so the SVM consistently mapped those types of flowsheet rows to the “Pain Rating 0-10 Scale” concept.

*Short description mapping will work in other domains.* This approach is generalizable to other short description mapping problems. Within the EHR, there are many types of data that are described with keyword-like phrases. For example, orders are typically created as a custom list for each healthcare organization. There currently is not a standardized list of orders, but even if one were to exist, the job of mapping each organizations’ orders to the standard would be time-consuming and error prone and would benefit from the approach described in this paper.

*There are some limitations with this research.* This approach was only tested using the Pain Information Model. The approach should work well on the other nine nursing IMs, but the work to carry out that validation still needs to be performed. A second limitation is that performance is dependent on the quality of the IM. Our goal was to map a flowsheet row to the best equivalent concept in the IM, which sometimes meant it was mapped to a higher level.
concept because an exact equivalent concept did not exist in the IM. Also, all of the organizations for this study use the same EHR vendor. It would be helpful to use this approach on flowsheet data types from other EHR vendors to ensure that the approach works with data from all vendors. Finally, the eight organizations involved in this research are diverse geographically, but tend to be medium to large in the size of their institutions and are all from inside United States (US). To fully validate this approach, flowsheet data should be obtained from smaller organizations and organizations from outside the US.

Crowd-sourced mapping. Additional work is needed to develop tools to make it easier to review the mapped data and to specify which labels are correct and which are not. This would make it easy for multiple reviewers from an organization to quickly review how the model labeled their data and also review mappings from other organizations. In this way, the “gold standard” training data would continue to get better and as new organizations join the workgroup, the predicted mappings from the automated process would also improve, thereby reducing the workload across the group.

Conclusion

This research shows that it is possible to use machine learning to automate the mapping of flowsheet rows to standard information models. Furthermore, the same approach should work wherever there is a need to map short description items to an information model. The model performance depends on having good training data from a variety of organizations. There is a need to develop tools and processes to support collaborative mapping of local EHR data to models so that the workload of doing the mapping can be distributed and the benefits of automating the mapping can be used by many organizations.

Acknowledgements

This work was supported in part by NIH NCATS grant UL1 TR002494. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

3. Observational Medical Outcomes Partnership (OMOP) [Internet]. [cited 2015 Jul 15]. Available from: http://omop.org/
Regional Variations in Documentation of Sexual Trauma Concepts in Electronic Medical Records in the United States Veterans Health Administration

Audrey L. Jones, PhD1,2, Warren B.P. Pettey, MPH, CPH1,2, Marjorie E. Carter, MSPH1,2, Emily Brignone, PhD1, Andrew Redd, PhD1,2, Ying Suo, MS1,2, Guy Divita, MS1, Rebecca K. Blais, PhD2,3, Jamison D. Fargo, PhD, MS2,3, Adi V. Gundlapalli, MD, PhD, MS1,2

1Informatics, Decision-Enhancement and Analytic Sciences (IDEAS) Center of Innovation, VA Salt Lake City Health Care System, Salt Lake City, UT; 2University of Utah School of Medicine, Salt Lake City, UT; 3Utah State University, Logan, UT

Abstract

Background: Experiences of sexual trauma are associated with adverse patient and health system outcomes, but are not systematically documented in electronic health records (EHR). Objective: To describe variations in how sexual trauma is documented in the Veterans Health Administration’s EHR. Methods: Sexual trauma concepts were extracted from from 362,559 clinical notes using a natural language processing pipeline. Results: We observed variations in the presence of sexual trauma in notes across five United States regions: Pacific, Continental, Midwest, North Atlantic, Southeast. We also observed variations in the types of notes used to document sexual trauma (e.g., mental health, primary care) and sources of sexual trauma (e.g., adult, childhood, military) mentioned in the EHR. Our findings illustrate potential differences in cultural norms related to patient disclosure of sensitive information, and provider documentation. Standardized protocol for eliciting and documenting sexual trauma histories are needed to ensure Veteran access to high quality, trauma-informed care.

Introduction

Experiences of sexual trauma are associated with adverse patient and health system outcomes and are important to elicit during patient-provider interactions. These experiences are, however, not routinely disclosed by trauma survivors in civilian1–2 and military settings3–5. Moreover, these adverse experiences are not systematically documented in patient medical records due a variety of reasons, including provider norms and patient preferences. Sexual trauma, which includes both experiences of assault and harassment, has been described as part of adverse childhood experiences, as well as in adulthood on college campuses, in the workplace, and in intimate partner relationships. We are increasingly aware of the detrimental clinical outcomes associated with sexual trauma. These include psychiatric comorbidities such as depression, anxiety, post-traumatic stress disorder, and alcohol and drug use disorders6–7. The initial psychiatric sequelae of sexual trauma, in the absence of intervention, may accumulate to affect a broad range of health and psychosocial domains, with symptoms often persisting for several years following the incident. Indeed, long-term consequences of sexual trauma include chronic medical conditions (e.g., chronic pulmonary disease, liver disease, and hypertension)8, legal involvement9, homelessness10, and suicide11.

Disclosure of a history of sexual trauma is an essential first step to receiving counseling or trauma-informed treatment that may address the long-term sequelae of these traumatic experiences12. However, little is known about the circumstances of sexual trauma disclosure in health care settings, or how patient histories of sexual trauma are documented in electronic health record (EHR) notes. Most health care systems lack routine screening for patient histories of sexual trauma. Even when screenings occur, sexual trauma experiences are under-reported due to stigma, shame, and privacy concerns13. For example, a recent study of documentation of sexual trauma in the Veterans Health Administration found that male and racial/ethnic minorities with a history of military sexual trauma were less likely than female and non-Hispanic Whites, respectively, to disclose information regarding sexual trauma on an initial clinical screen14. Some patients only volunteer sexual trauma
information in a therapeutic context, or when rapport has been established with their routine provider. In other cases, sexual trauma is documented when a patient presents to the emergency department for treatment of acute symptoms. An examination of the free text of EHR notes may shed light on the circumstances of patient disclosures of sexual trauma and provider practices of documentation in medical records.

Health care providers document a patient’s medical status, psychosocial circumstances, and historical information through clinical narratives in the EHR, and the presence and availability of this information in clinical documents informs patient care. In the case of sexual trauma, the availability of patient history information facilitates the delivery of trauma-informed care across the health care system\(^5\). However, there is reason to think that sexual trauma history information may not be captured in a standardized way. Use of International Classification of Disease (ICD) codes to record sexual trauma histories is limited, and research in other clinical domains has identified substantial variations in documentation practices\(^6\). To optimize health care efficiency, and to maximize the utility of the EHR, it is important to identify major variations in clinical documentation of sexual trauma histories and to understand the nature of those variations\(^6\).

This paper aims to characterize regional geographic variations in documentation of sexual trauma histories in the Veterans Health Administration (VHA). We chose United States geographic regions as a potential source of variation because these may serve as proxies for sociocultural norms in patient reporting and provider documentation of sexual trauma information\(^8\). To accomplish the study aim, mentions of sexual trauma concepts in the EHR were extracted via a natural language processing (NLP) pipeline. Prior studies have demonstrated the feasibility of using NLP techniques to identify mentions of sexual trauma in the EHR\(^17\), and have examined racial/ethnic and sex differences in patterns of sexual trauma disclosure\(^14\). To our knowledge, this is a first study to explore regional variations in sexual trauma documentation.

The VHA is an optimum setting to explore geographic variations in sexual trauma documentation because of the size of the healthcare system, and because patient histories of sexual trauma have been prioritized for documentation. Indeed, VHA is the largest integrated healthcare system in the United States, serving nearly 6 million unique patients every year through 152 medical centers and nearly 1000 outpatient clinics. Since 2004, VHA has implemented universal screening of all Veterans seeking care in its medical facilities for exposure to sexual trauma during military service (military sexual trauma, MST). Under this model, many patients with a history of sexual trauma have received access to specialized care designed to mitigate adverse outcomes associated with sexual harassment and assault. Eligible Veterans also receive treatment for experiences of sexual trauma occurring outside of military service (during childhood or adulthood). Information learned through an examination of regional variations in VHA medical records may apply to other healthcare systems in the United States, and to international health care settings where sexual trauma is prevalent\(^19\).

**Methods**

**Setting and Study Population**

This study examined documentation of sexual trauma concepts from a national, random sample of Veterans who served in recent conflicts as part of Operations Enduring Freedom/Iraqi Freedom/New Dawn (OEF/OIF/OND) and who received care in VHA medical facilities. We selected a random sample of 10,000 male and 10,000 female OEF/OIF/OND Veterans by linking a Department of Defense (DoD) roster of service members who served in the recent conflicts with administrative data stored in the VHA corporate data warehouse. Data were accessed through the VHA Informatics Computing Infrastructure (VINCI), a secure research portal\(^20\). Veterans were included in this study if they received care in VHA medical facilities, and had a valid MST screen result between October 2009 and October 2014. For each patient, we analyzed 12 months of electronic medical notes that are available in the VHA corporate data warehouse for use in research\(^10\). All study procedures were approved by the University of Utah Institutional Review Board and the Research Review Committee of the VA Salt Lake City Health Care System.

**Procedures**

We downloaded all electronic medical notes associated with outpatient visits for the study cohort in the 12 months following an initial MST screen result. We processed the 12-month set of electronic medical notes using an NLP pipeline, V3NP, that was developed to extract all sexual trauma concepts such as adverse childhood experiences and mentions of sexual trauma in adulthood (including trauma that occurred during military service) from VHA electronic medical records\(^17\). The pipeline was built on the APACHE-Unstructured Information Management applications (UIMA) framework\(^21\). In brief, the pipeline parses the electronic medical note into its component parts such as sections, content headings, lists, sentences, lines, and finally individual concepts (tokens). The individual tokens were then compared to a look-up dictionary of sexual trauma terms. The
dictionary was developed for this pipeline using terms from standard vocabularies and supplemented with expert opinion after review of a set of VHA electronic medical records as there was lack of complete coverage of sexual trauma concepts in the UMLS Metathesaurus and SNOMED-CT. An important feature of the pipeline was to include lexical variants of terms and to exclude concepts that were negated such as “no evidence of sexual trauma”. Thus, the goal of the pipeline was to extract “positively asserted” concepts. All concepts related to sexual trauma extracted by the NLP pipeline as positively asserted (with clinical text surrounding the extracted concept to provide context, “snippets”) were reviewed by a set of trained human reviewers for true positive mentions. As reported elsewhere, the overall positive predictive value for identifying sexual trauma at the individual concept level was 0.90\(^2\). The trained reviewers also determined the type of sexual trauma described in each mention: 1) childhood sexual trauma, 2) adult sexual trauma, 3) military sexual trauma (MST), 4) unspecified sexual trauma, or 5) no sexual trauma (i.e., false positive).

At the document level, we created a binary variable noting any evidence of sexual trauma (yes/no), and a categorical variable with specific type(s) of sexual trauma recorded: adult only, childhood only, MST only, unspecified only, or multiple types of trauma. Documents that only included false positive mentions of sexual trauma, and those documents lacking true positive mentions were coded as having no evidence of sexual trauma.

**Geographic Variables**

All clinical notes were affiliated with a VHA medical center or outpatient clinic where the patient received care. We extracted geocoded attributes of each facility from VHA administrative records. Facilities are coded as urban, rural, highly rural, or insular, based on rural-urban commuting area (RUCA) definitions using the standard VHA classification that has been validated for Veteran research\(^22,23\). We coded clinical notes as urban or non-urban, based on the RUCA classification (RUCA <=1.1 is urban, all other values are non-urban) of the facility where the note was written.

VHA facilities are also categorized into one of five geographic regions of the United States, based on aggregated Veterans Affairs Integrated Service Network (VISN) catchment areas: Pacific, Continental, Midwest, Southeast, and North Atlantic regions (Figure 1). We assigned clinical notes to one of the five geographic regions, based on the facility where the note was written.

![Figure 1: Five United States geographic regions. Region boundaries correspond to Veterans Affairs Integrated Service Network (VISN) catchment areas. Green circles indicate Veterans Affairs medical center locations.](image)

**Statistical Analyses**

We used cross-tabulations to characterize regional variations in notes containing sexual trauma documentation (yes/no), the type of notes mentioning sexual trauma, and types of sexual trauma mentions (e.g., military, adult, childhood, etc). We used logistic regressions to estimate regional differences in sexual trauma documentation (yes/no) and multinomial logistic regressions to estimate regional differences in the types of notes mentioning sexual trauma. All models controlled for patient sex, and used a robust standard error specification to account
for the clustering of notes within VHA patients. Wald joint tests of significance were used to simultaneously test region coefficients.

Results

Of 20,000 male and female VHA outpatients, 14,831 had at least one medical note during the study period. A total of 362,559 notes were processed from the study cohort using the NLP pipeline; and 4,178 of 4,937 (84.6%) notes contained true positive mentions of sexual trauma concepts. The NLP precision was slightly better in the Midwest region (note level positive predictive value [PPV]=88.4%), compared to the other United States regions: Pacific PPV=84.4%, Continental PPV=84.7%, Southeast PPV=82.8%, and North Atlantic PPV=83.0%. Regardless of sexual trauma mentions, the majority of all notes were observed in urban versus rural VHA facilities (n=324,053, 91.8%).

As shown in Table 1, there were regional variations in the documentation of sexual trauma (Wald $X^2 = 59.94$, $p<0.001$). The Southeast region had the largest average number of notes per patient, but a relatively low frequency of notes containing sexual trauma mentions. Only 15.6% of notes containing sexual trauma were in the Southeast region, compared to 20.4% of notes lacking sexual trauma mentions.

Table 1: Regional variation in number of notes, by presence of sexual trauma concept

<table>
<thead>
<tr>
<th>Geographic region</th>
<th>Notes lacking sexual trauma</th>
<th>Notes containing sexual trauma</th>
<th>Number of notes per patient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Pacific</td>
<td>58,775</td>
<td>16.4</td>
<td>745</td>
</tr>
<tr>
<td>Continental</td>
<td>73,045</td>
<td>20.4</td>
<td>1,005</td>
</tr>
<tr>
<td>Midwest</td>
<td>68,731</td>
<td>19.2</td>
<td>810</td>
</tr>
<tr>
<td>Southeast</td>
<td>72,933</td>
<td>20.4</td>
<td>651</td>
</tr>
<tr>
<td>North Atlantic</td>
<td>84,866</td>
<td>23.7</td>
<td>967</td>
</tr>
</tbody>
</table>

Notes Containing Sexual Trauma Concepts

Notes lacking sexual trauma concepts included 1,995 unique note titles, whereas notes containing sexual trauma concepts only included 365 unique titles. As shown in Table 2, notes lacking sexual trauma concepts were often addendums and administrative in nature; nursing, primary care, and mental health disciplines were also represented. In contrast to other notes, most of the notes containing sexual trauma mentions were mental health in nature. Smaller numbers of notes with sexual trauma mentions related to primary care or to VHA compensation and pension (C&P) examinations.

Table 2: Top 10 most frequent note titles, by presence of sexual trauma concept

<table>
<thead>
<tr>
<th>Titles of notes lacking sexual trauma concepts</th>
<th>N</th>
<th>%</th>
<th>Titles of notes containing sexual trauma concepts</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addendum</td>
<td>69,941</td>
<td>19.5</td>
<td>Mental health note</td>
<td>417</td>
<td>10.0</td>
</tr>
<tr>
<td>Administrative note</td>
<td>10,652</td>
<td>3.0</td>
<td>Psychology note</td>
<td>254</td>
<td>6.1</td>
</tr>
<tr>
<td>Nursing note</td>
<td>10,111</td>
<td>2.8</td>
<td>Psychiatry note</td>
<td>234</td>
<td>5.6</td>
</tr>
<tr>
<td>Mental health note</td>
<td>8,764</td>
<td>2.5</td>
<td>Addendum</td>
<td>144</td>
<td>3.5</td>
</tr>
<tr>
<td>Primary care note</td>
<td>8,316</td>
<td>2.3</td>
<td>Mental health outpatient note</td>
<td>137</td>
<td>3.3</td>
</tr>
<tr>
<td>Telephone encounter note</td>
<td>7,985</td>
<td>2.2</td>
<td>Mental health medication mgmt.</td>
<td>122</td>
<td>2.9</td>
</tr>
<tr>
<td>No show note</td>
<td>5,363</td>
<td>1.5</td>
<td>Group counseling note</td>
<td>118</td>
<td>2.8</td>
</tr>
<tr>
<td>Primary care nursing note</td>
<td>5,074</td>
<td>1.4</td>
<td>C &amp; P examination note</td>
<td>117</td>
<td>2.8</td>
</tr>
<tr>
<td>Primary care telephone note</td>
<td>4,637</td>
<td>1.3</td>
<td>Social work note</td>
<td>112</td>
<td>2.7</td>
</tr>
<tr>
<td>Medication management note</td>
<td>4,441</td>
<td>1.2</td>
<td>Primary care note</td>
<td>75</td>
<td>1.8</td>
</tr>
</tbody>
</table>
Classification of Note Titles

We reviewed the 365 unique note titles names, and combined note titles with similar names into groups. The note titles broadly fell into six main categories: primary care, mental health, social services, medical or surgical services, emergency or inpatient services, C&P exams, and other. The other category included labs, radiology, person-specific notes (e.g., physician, attending), addendums, and miscellaneous (e.g., no-show note, administrative note).

In notes lacking sexual trauma concepts (Figure 2a), the proportion of primary care and mental health notes were similar across regions of the United States. Of all geographic regions, facilities located in the Pacific states had a higher percentage of mental health notes than the other U.S. regions (15.1% vs. 12.1%-13.8% in other regions), and a smaller percentage of primary care notes (17.4% versus 19.2%-20.8% in other regions).

![Figure 2a: Types of notes lacking sexual trauma mentions, by geographic region](chart)

In contrast to notes lacking sexual trauma mentions, we observed greater regional variation in notes containing sexual trauma concepts (Figure 2b; Wald $\chi^2 = 176.99, p<0.001$). VHA facilities located in Pacific states had a greater percentage of primary care notes (28.3%), compared to facilities in the Southeast (13.1%) and other regions (17.3%-20.0%). Facilities in the North Atlantic had a higher percentage of emergency or inpatient service notes, compared to facilities in other regions (4.5% vs. 1.8%-3.7% in other regions).

![Figure 2b: Types of notes containing sexual trauma concepts, by geographic region](chart)
Region and Sex Variations in Sexual Trauma Mentions

We explored whether the patterns of regional differences in notes with sexual trauma concepts were related to patient sex. Most (94%) of all notes containing sexual trauma concepts belonged to female patients. Thus, the regional distribution of note titles among female Veterans was similar to the total distribution (males and females combined) presented in Figure 2b above. Among female patients, for instance, VHA facilities located in Pacific states were more likely than facilities in other regions to document sexual trauma in primary care (23.8%). VHA facilities in the Southeast were less likely than facilities in other United States regions to document sexual trauma in primary care (10.1%; data not shown). VHA facilities in Pacific states had lower percentages of sexual trauma notes from mental health (39.9%) compared to facilities in other regions (50.3%-60.3%; data not shown).

A different pattern was observed in the notes of male patients (n=248) than in notes of females. As shown in Figure 3, VHA facilities in the Continental U.S. region were more likely to document sexual trauma for male patients in primary care (19.6%), compared to facilities in other U.S. regions (7.7%-9.8%). VHA facilities in the Midwest and Atlantic regions were more likely than facilities in other regions to document sexual trauma concepts for male patients in mental health note types (75.4% and 70.7% versus 38.3%-59.1%). More than in other geographic regions, sexual trauma concepts in Southeast facilities were documented in social service (14.9%), medical/speciality (12.8%), and other (14.9%) notes. Tests of regional variation in the types of notes mentioning sexual trauma were statistically significant in the models for females and males (both Wald test p-values <0.001).

Types of Sexual Trauma Concepts: Geographic Variations by Sex

We also explored the types of sexual trauma concepts picked up by the NLP pipeline: military sexual trauma, sexual trauma occurring in adulthood (outside of the military), childhood sexual trauma, and other sexual trauma incidents where circumstances were not specified. The overall test of regional variations in type of sexual trauma recorded was statistically significant (Wald $X^2 = 99.95$, p<0.001), with distinct regional patterns in the type of sexual trauma documented for female versus male patients.
As shown in Figure 4a, sexual trauma notes for female patients in the Pacific states often referenced adult sexual trauma (43.2%), with relatively smaller percentages of sexual trauma concepts related to military service (34.1) or childhood (11.4%); 9.1% of sexual trauma notes in Pacific states referenced more than one type of trauma. For female patients, sexual trauma notes in the Continental and Midwest regions had the greatest percentage of childhood trauma references (27.5% and 27.7%, respectively); while sexual trauma notes in the Southeast and North Atlantic regions primarily related to military sexual trauma only (71.7% and 63.4%, respectively).

**Figure 4a:** Types of sexual trauma concept documented in notes of female patients, by geographic region

For male patients, the majority of sexual trauma notes referenced military sexual trauma only (56.0%-70.2% across regions; Figure 4b) or multiple types of trauma (9.6%-21.6% across domains). Regardless of region, male patients’ sexual trauma notes rarely referenced non-military sexual trauma.

**Figure 4b:** Types of sexual trauma concept documented in notes of female patients, by geographic region
Discussion

Patient histories of sexual trauma have important implications for patient and health system outcomes, but documentation of this sensitive and personal information lacks standardization. If sexual trauma information is not recorded at all, sexual trauma survivors may have more difficulty accessing high quality, trauma-informed care. In addition, lack of provider awareness of trauma histories could inadvertently lead to poor care delivery and re-traumatization of patients.

We sought to examine regional variations in the extent and nature of documentation of sexual trauma experiences in the VHA’s EHR. While there was relatively small geographic variation in the types of notes used to document non-sexual trauma phenomena, documentation of sexual trauma histories varied across the major United States regions. In the Pacific states, we observed relatively greater proportions of primary care notes and relatively smaller percentages of mental health notes, compared to the other geographic regions. We also found more variation in the types of sexual trauma documented in Pacific states, compared to other United States regions, particularly for female patients. It may be the case that primary care providers in Pacific states have received more training in eliciting sexual trauma histories, or experience less hesitation in recording this information in the EHR.

Our finding of unique patterns of results in male versus female patients highlights some areas where sexual trauma may not be reported due to fear of stigma. Our finding of relatively few mental health notes for male patients in the Southeast region, compared to other United States regions, suggests there may be increased barriers to seeking mental health care or to disclosing sexual trauma in the course of mental health treatment for men in the Southeast. In the Southeast region, we also observed relatively small percentages of sexual trauma notes in primary care and relatively small percentages of non-military types of sexual trauma documented, compared to other United States regions. Importantly, the performance of the NLP pipeline was similar across geographic regions. While the current study was not designed to explain reasons for lack of documentation, our results suggest that cultural factors may be at play. For example, the Southeast is thought to be more culturally conservative than other United States regions, and this could explain relatively low rates of disclosure and documentation of sexual trauma in the clinical setting. Future research is warranted to elucidate how patients and providers navigate discussions of sexual trauma in diverse clinical settings in the Southeast and in other United States regions.

We note study limitations. First, this study did not systematically evaluate false negative captures of military sexual trauma. Additional work is evaluate to determine the performance of the NLP pipeline overall, and across diverse cultural contexts. Second, the study sample included small numbers of notes for male patients in each geographic region, which precluded multivariable analyses that might help to explain regional patterns. Specifically, the study did not examine patient, provider, or health system factors that may contribute to variations in sexual trauma documentation. It is possible that some of the regional variations observed in the current study could be due to patient differences in sexual trauma exposure and help-seeking or to patients requesting that their trauma not be documented. Despite limitations, this descriptive analysis illustrates the utility of NLP methods to detect systematic differences in the documentation of sensitive and personal information in the EHR. Our future work will consider provider characteristics (e.g. age, gender) in combination with patient factors to better understand the variations in sexual trauma documentation observed here.

Conclusions

While there is increasing public health attention on sexual harassment and assault, little is known about how sexual trauma histories are documented and managed in integrated health care systems. To fill this knowledge gap, this study examined the clinical context of sexual trauma documentation, and the nature of traumas reported. We found variations in documentation that are unlikely to be explained by differential exposures alone. Standardized mechanisms of collecting and documenting sexual trauma histories are needed to ensure Veterans’ access to high quality, trauma-informed care. Provider training designed to increase awareness of the cultural barriers and challenges of eliciting and documenting sexual trauma in clinical notes may be an important first step.

Acknowledgements

Funding for this project was provided by grant IIR 12-084 (PI: A.V.G.), and VA Center of Innovation Award #I50HX001240 from the Health Services Research and Development of the Office of Research and Development of the US Department of Veterans Affairs. Dr. Jones is supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002538 and
The views expressed in this article are those of the authors and do not necessarily represent the position or policy of the U.S. Department of Veterans Affairs, the National Institutes of Health, or the U.S. government.

References


Abstract

Transition words add important information and are useful for increasing text comprehension for readers. Our goal is to automatically detect transition words in the medical domain. We introduce a new dataset for identifying transition words categorized into 16 different types with occurrences in adjacent sentence pairs in medical texts from English and Spanish Wikipedia (70K and 27K examples, respectively). We provide classification results using a feedforward neural network with word embedding features. Overall, we detect the need for a transition word with 78% accuracy in English and 84% in Spanish. For individual transition word categories, performance varies widely and is not related to either the number of training examples or the number of transition words in the category. The best accuracy in English was for Examplification words (82%) and in Spanish for Contrast words (96%).

Introduction

Even with the increased use of video and multi-media, text remains an important tool for patient education. However, creating well-written, understandable text is difficult and requires appropriate vocabulary use, correct grammar, and good writing style. One important component for well-written text is the use of transition words to help the reader follow the flow of ideas.

Transition words (also referred to as connectives) are words or phrases that connect linguistic units in a text. For example, single term transition words include “however, then, also, but, thus, so, therefore, and, still, furthermore, rather, or, meanwhile”, and multi-term examples include “even so, in conclusion, other than, as a result, in addition, in the meantime”. In this work, we will use the phrase “transition word” to refer to both single and multi-term transition words. Transition words are categorized according to their role, e.g., “initiating a topic, adversative, comparison, results”. Several different categorizations have been proposed depending on the subfield or goal. In this work, we use 16 categories paralleled across English and Spanish. Table 1 gives a list of the different category types and including examples in English.

Transitions words play an important role in text flow, especially for introducing a new topic for a reader to learn about. The words provide logical connections, help convey the structure of an argument, show information flow, improve coherence and, because of these, affect reader comprehension. Additionally, their use can be an indicator of text type. Several studies have found that their frequency differs depending on the type of text, such as news, conversation, fiction, or academic prose, with the latter containing the most transition words.

In this paper, we focus on the task of predicting transition words in medical text. In general, transition words can occur at the beginning of a sentence or embedded in a sentence to link clauses. We examine the former, i.e. predicting transition words at the beginning of a sentence. For example, given the pair of sentences

After the second year of growth hormone therapy, beneficial bone growth decreases. ____ GH therapy is not a satisfactory long term treatment.

we aim to predict that a transition word should occur at the beginning of the second sentence where the underscore is, specifically, a Results transition category word. In the text that this example was derived from, ‘therefore’ was used, however, other transition words from the same category could have also been used, e.g. ‘thus’ or ‘as a result’. We pose two variants of this prediction task. First, predict if any transition word should be used at the beginning of a sentence. Second, predict if a particular category of transition word should be used.

Previous work has shown how transition words signal relationships between and among the idea hierarchy in text using manual approaches. We view the problem as a supervised learning problem. We introduce a new dataset for transition word prediction in medical texts with data in two languages, English and Spanish. Using this dataset, we establish a baseline using a feedforward neural network trained using backpropogation and provide initial analyses on the difficulty of predicting the individual categories. We conclude with a number of possible future directions.
Historically, readability formulas were popular for guiding simplification and have been used to replace difficult terms. However, these formulas often require large parallel corpora, particularly in the medical domain, and are not helpful as writing guidelines. They also do not leverage modern resources that are currently available, e.g., large corpora with term frequencies to suggest terms with high familiarity as a replacement for difficult terms.

In recent years, computational approaches to improving the readability and understandability of health-related text have started to dominate. Several approaches have been tested as part of rule-based simplification algorithms. For example, when simplifying single words, term frequency (often measured by the Google Web Corpus) is used to replace

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of words</th>
<th>Example English connector words in category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement</td>
<td>35</td>
<td>again, in the first place, not only, similarly, as a matter of fact granted that, as long as, so long as, on condition that</td>
</tr>
<tr>
<td>Cause</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Conclusion &amp; Consequence</td>
<td>53</td>
<td>in conclusion, in brief, in summary, to sum up, all in all</td>
</tr>
<tr>
<td>Contrast</td>
<td>11</td>
<td>however, in contrast, nevertheless, yet, on the other hand</td>
</tr>
<tr>
<td>Emphasis</td>
<td>4</td>
<td>as a matter of fact, in fact, actually, indeed</td>
</tr>
<tr>
<td>Exemplification</td>
<td>55</td>
<td>for example, for instance, to illustrate, notably, in general</td>
</tr>
<tr>
<td>Explanation</td>
<td>5</td>
<td>that is to say, that is, namely, in other words, put differently</td>
</tr>
<tr>
<td>Importance &amp; Order</td>
<td>6</td>
<td>most importantly, primarily, above all, most significantly</td>
</tr>
<tr>
<td>Linking</td>
<td>7</td>
<td>as for, with respect to, regarding, with regard to, as far as</td>
</tr>
<tr>
<td>Location</td>
<td>38</td>
<td>in the middle, to the left, alongside, behind, across,</td>
</tr>
<tr>
<td>Opposition</td>
<td>38</td>
<td>different from, of course, but, even so, then again, while</td>
</tr>
<tr>
<td>Particularization</td>
<td>4</td>
<td>in particular, particularly, more specifically, specifically</td>
</tr>
<tr>
<td>Results</td>
<td>5</td>
<td>as a result, as a consequence, therefore, thus, accordingly</td>
</tr>
<tr>
<td>Sequence</td>
<td>22</td>
<td>first, first of all, to begin with, for one thing, moreover</td>
</tr>
<tr>
<td>Similarity</td>
<td>5</td>
<td>likewise, similarly, correspondingly, in the same way, also</td>
</tr>
<tr>
<td>Time</td>
<td>63</td>
<td>at first, then, afterwards, later, to begin with, meanwhile</td>
</tr>
<tr>
<td>Total</td>
<td>387</td>
<td>423</td>
</tr>
</tbody>
</table>

Table 1: The number of transition words for each of the 16 categories in English and Spanish along with example transition words in English.

Background

We focus on transition words in the context of text simplification of medical text. The goal of text simplification is to transform texts into simpler variants that are easier to understand by a broader audience without losing content. Our project’s goal is to accomplish this transformation by providing a text editor that guides writers in the simplification process. The tool is developed using data-driven algorithms that are trained on large sets of text and evaluated in user studies to show evidence-based outcomes. Since transition words play an important role in text comprehension, e.g., it has been shown that not understanding them limits text comprehension, making the transition between sentences explicit may assist readers with text comprehension since it helps clarify the logical connections in a text.

Simplifying text to increase reading comprehension for medical and healthcare information is an important task since it may help increase health literacy of patients and their family. Each year, chronic diseases afflict more people and treatments are becoming increasingly more complex, often requiring the need for participatory medicine where patients take an active role in their healthcare. However, the simplification task is difficult because of the complexity of the topics; the lack of large parallel corpora, particularly in the medical domain; and the need for simplifications that do not omit important information or introduce errors.

Historically, readability formulas were popular for guiding simplification and a large, if somewhat older, non-computational group of projects focused on applying readability formulas such as the Flesch-Kincaid grade level formula. The formulas are easily accessible since they are available in common text editing software (e.g., Microsoft Word), online (e.g., http://www.readabilityformulas.com), or through commercial avenues. This research trend is declining since using readability formulas has not been successful. Studies showing a positive effect of their use on reader comprehension are rare. Readability formulas do not identify what aspects of a text are difficult, do not provide writing alternatives, and use features that are rudimentary, e.g., equate word length with difficulty. These formulas do not incorporate current knowledge about the reading process, have limited scientific basis, and are not helpful as writing guidelines. They also do not leverage modern resources that are currently available, e.g., large corpora with term frequencies to suggest terms with high familiarity as a replacement for difficult terms.

In recent years, computational approaches to improving the readability and understandability of health-related text have started to dominate. Several approaches have been tested as part of rule-based simplification algorithms. For example, when simplifying single words, term frequency (often measured by the Google Web Corpus) is used to replace
terms with easier synonyms\(^1\). Several lexical simplification approaches have been suggested that use either hand-crafted rewrite rules\(^2\) or learned rules\(^3\).

In addition to rule-based simplification, there is also a health research stream using machine learning approaches. The algorithms vary from traditional machine learning to evaluate different features\(^4\) to newer deep learning neural networks trained on parallel corpora\(^5\). Evaluations vary, including descriptive comments on outcomes\(^6\), subjective evaluations by experts or laypersons\(^7\), and user studies measuring impact through reading comprehension tasks\(^8\).

While good progress has been made recently on text simplification algorithms, particularly with data-driven approaches, the output is still not good enough quality to be used in real-world applications. Instead, for our project, we focus on human assisted simplification by using automated approaches to help guide content creators in simplifying medical text. For transition words, we plan to identify possible locations in the text where transition words could be used and suggest the category of the transition word, along with a ranked set of candidate transition word options. This paper is the first step towards this goal by evaluating the performance of supervised approaches to the transition word prediction problem.

**Dataset creation**

We created two datasets for the transition word prediction task: one in English and one in Spanish. English and Spanish are the two most commonly spoken languages in the U.S. and examining the task in two languages allowed us to compare how prediction performance differs across languages. Both datasets were based on medical Wikipedia articles using a large list of transition words. Both the English and Spanish datasets are publicly available online along with the transition word lists used to create them.

**Corpora**

Since the frequency of transition words differs depending on the topic of the corpus, it is important to train on a representative corpus. We aim to simplify health educational text and created a corpus from general information on common diseases. We downloaded all articles in English and Spanish Wikipedia tagged with the category “Diseases and disorders”. The English corpus contains 636 articles for a total of 1.3M words. The Spanish corpus contained 493 articles for a total of 637K words. We used these two corpora to derive our transition word examples.

**English Dataset**

To generate our transition word prediction datasets, we first identified a list of transition words in English from different online educational and writing sites listing transition words. We combined information and created 16 categories of transition words containing 387 transition words. Table 1 gives an overview of the categories with the number of transition words in both languages in the category along with example English words for each category. The categories range in size from as large as 63 words (Time) to as small as 4 (Emphasis and Particularization).

To generate the classification examples, we considered every pair of adjacent sentences within the corpus. If the second sentence in a pair contained one of the transition words, then it was marked as a positive example, otherwise, it was marked as a negative example, i.e. when there was no a transition word between the two sentences. This results in a dataset of positive examples, \(<s, s, c>\), and negative examples, \(<s, s>\), where \(s\) is a sentence, \(s\) is the sentence following \(s\), and \(c\) is one of the connectors.

Table 2 shows the total number of examples extracted from the corpus. Overall, just under a quarter of the sentence pairs had connectors in them and were marked as positive. We then further processed the dataset to ensure that each sentence in each example contained at least one noun. This led to the removal of a few inappropriate pairs and a final set of more than 70K examples.

**Spanish Dataset**

Guided by the English transition word list, we had a native, expert Spanish speaker generate a comparable list for Spanish. In most cases, there was a corresponding phrase in Spanish, however, there were a few exceptions. In some cases, this resulted in a reduction of the number of words, for example when there was not a good equivalent in Spanish or if several English transition words corresponded to the same one in Spanish. In other cases, there were multiple equivalent Spanish words for a single English word. The final Spanish transition word list contained the same 16 categories with 423 transition words, slightly more than the English equivalent.

---

\(^1\) https://github.com/dkauchak/ConnectorDataset2019

525
Similar to the English dataset, we generated the dataset in Spanish by considering all consecutive sentences in our corpus and creating positive, i.e., those containing a transition word in the first position of the second sentence, and negative examples. Because there are fewer texts available in Spanish Wikipedia covering the same topics, our Spanish corpus is smaller as is the number of positive and negative examples. The sentence pairs were then further preprocessed in a similar manner as English to ensure each sentence contained at least one noun. The resulting dataset contains approximately 27K sentence pairs (Table 2).

### Predicting Transition Words

Given a pair of adjacent sentences, the goal is to predict whether a transition word should be used between the sentences, specifically, at the beginning of the second sentence. We view this problem as a supervised classification problem and use a feedforward neural network trained on examples of sentence pairs to make both the binary prediction of whether a transition word should be used or not, as well as a more fine-grained prediction for each of the categories of transition words.

#### Features

Transition words are used to indicate how information in the first sentence relates to information in the second sentence. To capture this, we extracted features based on the nouns in the two sentences, which are often the content-bearing words in the sentence. We selected the first five nouns in each of the sentences ($s_1$ and $s_2$) using the Stanford CoreNLP toolkit. For each noun, to generalize beyond its lexical form, we then extracted a 300-dimension word embedding. The embeddings were acquired from large publicly available sets using the FastText pre-trained model. We chose this particular embedding model since it is publicly available, it is based on a large corpus, and it is available in both English and Spanish.

Based on these word embeddings, we extracted 3,601 features:

- The 300-dimension word embedding for each noun (300 for each of the 5 nouns in $s_1$ and the 5 nouns in $s_2$ for a total of 3,000 features). If a sentence contained less than five nouns, those features were given an empty value.
- The average vector of the noun word embeddings per sentence (300 features for $s_1$ and 300 features for $s_2$).
- The cosine similarity between the average word embeddings of the sentence pairs (1 feature).

#### Classifier

Since the task is new, we chose to use a classifier with well-known properties and used a feedforward neural network. We used the R RSNNS package with backpropagation learning, a 0.1 learning rate, and 1 hidden layer. To try and avoid overfitting, we tuned the number of hidden nodes and the number of training iterations on a development set. To pick the number of hidden nodes, we trained the network using 5, 10, and 50 hidden nodes and picked the best performing model based on the accuracy on the development set. We picked the number of training iterations as the

---

### Table 2: Dataset Description

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia Corpus</td>
<td>636</td>
<td>493</td>
</tr>
<tr>
<td>Sentence Pairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Example</td>
<td>16,315</td>
<td>7,178</td>
</tr>
<tr>
<td>Negative Examples</td>
<td>53,938</td>
<td>20,466</td>
</tr>
<tr>
<td>Total</td>
<td>70,253</td>
<td>27,644</td>
</tr>
<tr>
<td>Prepared Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Example</td>
<td>16,285</td>
<td>7,140</td>
</tr>
<tr>
<td>Negative Examples</td>
<td>53,859</td>
<td>20,290</td>
</tr>
<tr>
<td>Total</td>
<td>70,144</td>
<td>27,430</td>
</tr>
</tbody>
</table>
point where the standard squared error started to increase on the development set. Parameter selection was done per task.

**Experiments**

We examined two prediction tasks in both English and Spanish. The first task is to predict whether any connector should be used (Any-vs-None), i.e. the binary prediction between those examples that have a connector and those that don’t. The second task is, for each connector type, to predict whether that connector type should occur or no connector should occur, i.e. 16 separate binary prediction tasks.

**Experimental Setup**

To create a more balanced dataset, for each task, we downsampled the number of negative examples so that the proportion of examples was 1/3rd positive and 2/3rds negative, resulting in a majority baseline of 66.67% by always predicting no connector. For the Any-vs-None task, this resulted in 16,285 positive and 32,570 negative examples for the English task and 7,140 positive and 14,280 negative examples for the Spanish task. For the per category prediction task, we did not classify any category that had fewer than 100 positive examples. Tables 4 and 5 show the number of examples for each of the categories in English and Spanish, respectively. There were five categories in English and five in Spanish with fewer than 100 examples that were ignored.

We used 10-fold cross validation for all experiments and randomly selected one fold of the training portion as the development set for parameter tuning during training. We evaluated the models using accuracy, precision, and recall.

**Any-vs-None Classification**

Table 3 shows the results for the Any-vs-None classification task. In both English and Spanish, the accuracy is well above the baseline of 66.67%. This is particularly encouraging given how many different transition words there are in this task: sentence pairs that benefit from transition words do have marked differences from those that do not use them. Across all metrics, performance was better in Spanish than in English, even though the Spanish dataset had fewer examples and the number of individual transition words was larger. In both languages, the precision and recall numbers were similar indicating that there wasn’t a strong bias of the classifier towards either the positive or negative class. Given how varied the transition words are and the

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>77.97</td>
<td>66.21</td>
<td>69.27</td>
</tr>
<tr>
<td>Spanish</td>
<td>83.57</td>
<td>75.09</td>
<td>76.29</td>
</tr>
</tbody>
</table>

Table 3: Any-vs-None classification results for the two languages.

**Per Category Classification**

Tables 4 and 5 show the results for individual transition categories for English and Spanish, respectively.

**Accuracy:** Overall, the accuracy for most categories is lower than for the All-vs-None results, with the exception of the Exemplification category in English and the Contrast and Oppositions categories in Spanish. In a few categories, the classifier was unable to perform better than the majority baseline (two in English and four in Spanish). The per category transition prediction problem is generally much more difficult than just predicting whether a transition word should be used or not.

To try and understand the cause of the variation in performance per category, we calculated a one-tailed Pearson correlation between accuracy and the number of examples, the number of transition words in a category, and the number of examples per transition category. There were no significant correlations found for accuracy either in English or Spanish:
The number of examples is not related to accuracy, i.e., having a larger data set (since we keep the ratio between positive and negative examples identical) did not benefit the classifier.

The number of possible transition words was not correlated to accuracy, i.e., having a wider variety of positive examples did not correlate with accuracy.

The average number of examples per transition category (since the number of transition words varies for the different categories) did not correlate with accuracy, i.e., having more examples per individual transition category did not matter.

**Precision**: Generally, precision per category was around 50-60% for many of the categories in both languages, though there are some exceptions, particularly the best performing categories in the two languages. The lowest precision was the same category in both English (35%) and Spanish (41%), *Similarity*.

Similar to accuracy, we calculated the correlation between precision and the same three dataset metrics for both English and Spanish. In this case, we found one significant positive correlation between precision and the number of different transitions words in a category ($r = 0.555, p = 0.038$) for the English data, i.e. precision was better for transition list categories with more words in them. However, taking Bonferroni correction for performing multiple statistical tests, the correlation would not be considered significant. There were no significant correlations for the Spanish data.

**Recall**: Overall, recall was the lowest of the three evaluation metrics. This hints at the complexity of the problem: identifying all of the occurrences of transition word usage is difficult. English, in particular, had several categories with very poor recall scores: *Similarity* (12%), *Cause* (19%), *Contract* (20%), *Time* (20%) and *Opposition* (19%). Similar to precision, the lowest performing category for recall in both languages was *Similarity*.

There were no significant correlations found for English or Spanish between recall and the number of examples, number of transitions words, or number of examples per transition word.

<table>
<thead>
<tr>
<th>Transition Word</th>
<th>Ex. (Ct)</th>
<th>A (%)</th>
<th>P (%)</th>
<th>R (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement</td>
<td>2,055</td>
<td>69.89</td>
<td>60.11</td>
<td>31.24</td>
</tr>
<tr>
<td>Cause</td>
<td>2,885</td>
<td>67.11</td>
<td>52.65</td>
<td>18.82</td>
</tr>
<tr>
<td>Conc. &amp; Cons.</td>
<td>1,377</td>
<td>73.64</td>
<td>67.94</td>
<td>40.81</td>
</tr>
<tr>
<td>Contrast</td>
<td>1,202</td>
<td>67.83</td>
<td>57.13</td>
<td>20.46</td>
</tr>
<tr>
<td>Emphasis</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exemplification</td>
<td>671</td>
<td>81.52</td>
<td>78.85</td>
<td>62.28</td>
</tr>
<tr>
<td>Explanation</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imp. &amp; Order</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linking</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>1,984</td>
<td>68.16</td>
<td>56.04</td>
<td>29.13</td>
</tr>
<tr>
<td>Opposition</td>
<td>2,488</td>
<td>66.53</td>
<td>55.09</td>
<td>18.85</td>
</tr>
<tr>
<td>Particularization</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>301</td>
<td>72.55</td>
<td>63.74</td>
<td>43.87</td>
</tr>
<tr>
<td>Sequence</td>
<td>703</td>
<td>74.77</td>
<td>66.94</td>
<td>49.48</td>
</tr>
<tr>
<td>Similarity</td>
<td>224</td>
<td>64.29</td>
<td>34.69</td>
<td>11.60</td>
</tr>
<tr>
<td>Time</td>
<td>2,222</td>
<td>67.18</td>
<td>54.75</td>
<td>19.67</td>
</tr>
</tbody>
</table>

**Table 4**: English transition words classification per category (Ex = Example count, A = Accuracy, P = Precision and R = Recall).

<table>
<thead>
<tr>
<th>Transition Word</th>
<th>Ex. (Ct)</th>
<th>A (%)</th>
<th>P (%)</th>
<th>R (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement</td>
<td>867</td>
<td>63.94</td>
<td>44.16</td>
<td>26.76</td>
</tr>
<tr>
<td>Cause</td>
<td>1,647</td>
<td>68.73</td>
<td>57.55</td>
<td>31.33</td>
</tr>
<tr>
<td>Conc. &amp; Cons.</td>
<td>360</td>
<td>67.22</td>
<td>52.95</td>
<td>31.39</td>
</tr>
<tr>
<td>Contrast</td>
<td>524</td>
<td>96.50</td>
<td>95.24</td>
<td>94.27</td>
</tr>
<tr>
<td>Emphasis</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exemplification</td>
<td>250</td>
<td>80.40</td>
<td>72.20</td>
<td>67.20</td>
</tr>
<tr>
<td>Explanation</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imp. &amp; Order</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linking</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>350</td>
<td>66.00</td>
<td>51.59</td>
<td>21.43</td>
</tr>
<tr>
<td>Opposition</td>
<td>983</td>
<td>84.91</td>
<td>85.58</td>
<td>66.93</td>
</tr>
<tr>
<td>Particularization</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence</td>
<td>650</td>
<td>62.31</td>
<td>43.32</td>
<td>20.62</td>
</tr>
<tr>
<td>Similarity</td>
<td>404</td>
<td>64.60</td>
<td>41.94</td>
<td>17.05</td>
</tr>
<tr>
<td>Time</td>
<td>865</td>
<td>73.64</td>
<td>63.86</td>
<td>54.34</td>
</tr>
</tbody>
</table>

**Table 5**: Spanish transition words classification per category (Ex = Example count, A = Accuracy, P = Precision and R = Recall).
Discussion

In this paper, we have introduced the new problem of predicting transition word usage between adjacent sentence pairs. We created a dataset consisting of medical text examples divided into 16 categories of transition words in both English and Spanish, which is publicly available. Using this dataset, we examined two classification tasks: predicting whether any transition word should be used and predicting the category of the transition word. As a first approach, we used a feedforward neural network with word embedding features trained using backpropagation. We found strong results for the general transition word prediction task (English accuracy of 78% and Spanish of 84%, over a majority baseline of 67%), but more sporadic results for predicting individual connectors with accuracies as high as 97% and as low as 64%.

There are a number of possible directions for future research using this dataset. We viewed the prediction of each category as an independent binary prediction task. This allowed us to understand the difficulty of the transition word prediction task per category and to compare across English and Spanish. For practical applications, this problem should be viewed as a multi-class classification problem where the goal is to predict which of the 16 categories should be used, combined with the general prediction task (either as a separate step or as part of the multi-class prediction problem). Even given the category, there is still the question of which transition word to use. For use in a simplification tool, this then becomes a ranking problem among the possible transition words within a category.

As a first pass to aid analysis, we made some simplifying assumptions in the classification setup that should also be explored. First, we downsampled the negative examples to create a balanced dataset across all examples. For real application, this imbalance will have to be handled appropriately by the classifier. In particular, more evaluation needs to be done to examine how well the classifiers perform when applied to real texts where predictions must be made between every pair of sentences. Second, we ignored predicting on five of the transition categories with less than 100 examples. More data would likely fix this problem, particularly if the dataset were expanded beyond medical-related text.

The per category prediction results were low for many categories. Correlation analysis showed that this was the case regardless of the number of examples available or the number of transition words in the category. Recall, in particular, was generally low across all categories, suggesting that more work needs to be done to identify features that can better differentiate the broad usage of connector words. As an example, Exemplification had consistently strong results. We speculate that the performance was better than other categories because of the nature of the notion of exemplification which is perhaps less abstract than others (i.e., cause) and followed by more predictable and more local structure (an enumeration, rephrasing or repetition) than others. Other categories, however, may require more information for prediction.

Finally, in linguistics and education, the importance of transition words is well established and recognized, however, they have received little attention in the medical informatics and natural language processing communities and we believe they can be useful for a variety of tasks. For example, they can help with summarization by hinting at introductions, ordering, contrast and conclusions. For fact checking and information extraction tasks they can be used to identify and add structure to unstructured information sources. In general, transition words provide guidance about the flow of ideas and can be leveraged by many discourse-level tasks. We hope that the dataset, including the categorized list of transition words, along with the initial classification attempts helps motivate such uses.

Acknowledgements

Research reported in this paper was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


A Potential Answer to the Alert Override Riddle: Using Patient Attributes to Predict False Positive Alerts

Timothy I. Kennell Jr., BS and James J. Cimino, MD
Informatics Institute, University of Alabama at Birmingham, Birmingham, AL

Abstract

Electronic health records (EHRs) use alerts to help prevent medical errors, yet clinicians override many of these alerts due to desensitization from constant exposure (alert fatigue). We hypothesize that a clinician might override an alert warning about the dangers of a treatment if the patient’s health is so poor that the treatment is worth the risk or if a patient’s health suggests the treatment is not needed. We used logistic regression with general estimating equations to determine if the Early Warning Score (EWS), a measurement used to predict critical care need, could be used to predict alert overrides. EWS was a significant predictor of overrides for three alerts. Although EWS could not predict overrides for all alert rules, these results suggest that EWS may be helpful for some alerts, but that additional EHR data will be needed for predicting override behavior to a useful degree.

Background and Significance

A hallmark of electronic health record (EHR) systems is the implementation of logic to issue automated alerts and reminders as a form clinical decision support (CDS). Many of these alerts provide clinicians two options: an override option and a non-override option (typically an order cancellation). While this type of CDS can provide useful information when relevant, the messages it provides (referred to in this paper as “alerts”) are often judged as false positives by the receiving clinicians, leading to high override percentages.1 One study showed that primary care physicians can receive up to 56 alerts per day just from an email notification system2 and several studies point to override percentages in the range of 49 – 96%.13,4 Excessive override rates are problematic because they disrupt workflow, reduce the clinician’s quality of life, and can produce “alert fatigue” that may result in “pseudo-false positive” alerts that are inappropriately overridden, to the possible detriment of the patient.5

Methods of reducing the alert volume and the override rate typically focus on improving alert logic to increase its specificity (that is, replace false positives with true negatives).1 Such approaches are complicated by the fact that clinicians do not always accurately report their reasons for overrides.6 Other approaches have included manual review of the literature to prioritize alerts,7 maintenance of a drug-drug interaction database containing alert priorities,8 and dashboard construction, allowing easier manual intervention.9 One study proposed an automated method for predicting overrides per clinician based on previous actions by the specific provider in question and suppressing such alerts going forward.10 Each of these methods is either time-consumptive or not patient-specific.

While alerts typically use basic criteria to determine whether an alert should fire, studies indicate that physicians regularly consider more than what the EHR is using.11 We presume that a significant number of alerts are overridden, not because of fatigue, disagreement with the alert logic, or general stubbornness, but because the clinician has some awareness of the patient’s particular situation that is not covered by the logic. Indeed, although an override of an alert is not the same as a false positive alert, studies have shown many alert overrides are appropriate.12 We suspect that information about the situation may be available in the EHR at the time of the alert and could be used to suppress alerts that are predicted to be overridden. To our knowledge, the use of just-in-time patient-specific data outside that used in the alert logic itself has not been studied. Despite the limited criteria used by alerts, the EHR is overflowing with additional information that might reflect a physician’s decision when overriding an alert. One such example might be a patient’s overall health status. For example, a clinician might ignore a reminder to administer an influenza vaccine if the patient is currently undergoing cardiopulmonary resuscitation. Such an overall health status of a patient is readily available in most EHRs called the Early Warning Score (EWS) (see Table 1 for details on the score calculation.). The EWS is an aggregate score calculated from a patient’s vital signs each time they are recorded and applies increasing values to a patient’s worsening vital signs in order to identify patients at risk of deteriorating.13,14 The score is used for both prognosis of disease and indication for intensive care admission.13
**Table 1.** Early Warning Score Calculation. The column on the left indicates the parameter measured and the remaining column headers indicate the points added based on the range that the parameter falls into. Table format from Subbe, et. al.\(^1\)

<table>
<thead>
<tr>
<th><strong>Physiological Parameter</strong></th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic Blood Pressures</td>
<td>≤ 90</td>
<td>91–100</td>
<td>101–110</td>
<td>111–219</td>
<td></td>
<td></td>
<td>≥ 220</td>
</tr>
<tr>
<td>Temperature in °F</td>
<td>≤ 95</td>
<td>95.1–96.8</td>
<td>96.9–100.4</td>
<td>100.5 - 102.2</td>
<td></td>
<td></td>
<td>≥ 102.3</td>
</tr>
<tr>
<td>Heart Rate</td>
<td>≤ 40</td>
<td>41–50</td>
<td>51–90</td>
<td>91–110</td>
<td>111–130</td>
<td></td>
<td>≥ 131</td>
</tr>
<tr>
<td>Respiratory Rate</td>
<td>≤ 8</td>
<td>9–11</td>
<td>12–20</td>
<td></td>
<td>21–24</td>
<td></td>
<td>≥ 25</td>
</tr>
<tr>
<td>Oxygen Saturation</td>
<td>≤ 91</td>
<td>92–93</td>
<td>94–95</td>
<td></td>
<td>≥ 95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oxygen Therapy</td>
<td>Responses not equal to Room air or CPAP adds 2 points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level of Consciousness</td>
<td>Not Awake, alert; Awake, sedated; Obey commands; Responds to pain; Sleeping, difficult to arouse; Sleeping, easy to arouse; or Unable to assess adds 3 points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Objective**

We reason that extenuating information present in the record, such as the EWS, could be used to suppress irrelevant alerts, even when the alert logic does not consider it. Furthermore, we hypothesize that clinicians override alerts, in part, because of something they know about the patient that is not represented in the alert logic. This paper describes a study of the correlation of patient characteristics *unrelated to specific alert logic that is nevertheless present in the EHR*, with the intent of using such correlations to suppress alerts that are likely to be overridden anyway, thus reducing their adverse effects. We report the application of the above approach with the use of a general measure of patient condition severity.

**Methods**

**Data Source**

All data used in the study came from the University of Alabama at Birmingham’s (UAB) University Hospital. Specifically, the information came from both the Cerner Millenium EHR and PowerInsight, a research copy of the EHR’s database. All data were acquired with IRB approval for data reuse with waiver of consent.

**Alerts**

Initially, we acquired all alerts events ever triggered in the EHR from the time of its implementation in 2010 to 2017. Events that resulted from alerts having no override option were removed from the set as they did not pertain to the objective of the study. The remaining alerts are listed in **Table 2** with a brief description to supplement the name of the alert. Despite the apparent duplication of some alerts, these function in different, nuanced capacities in UAB’s EHR.

**Table 2.** Descriptions of alerts used in the analysis of alert overrides and the early warning score.

<table>
<thead>
<tr>
<th><strong>Alert Name</strong></th>
<th><strong>Alert Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLERGYDRUG</td>
<td>Patient allergic to ordered medication</td>
</tr>
<tr>
<td>AMB_RULE10_FEC_SPEC_ALERT</td>
<td>Order requires a fecal specimen not yet collected</td>
</tr>
<tr>
<td>AMB_RULE11_TIMED_UR_ALERT</td>
<td>Order requires a urine sample not yet collected</td>
</tr>
<tr>
<td>AMB_RULE30_LUNG_CANCER</td>
<td>Prior, negative lung cancer CT screen within 24 hrs.</td>
</tr>
<tr>
<td>AMB_RULE8_FIT_TEST_ALERT</td>
<td>Fecal iron test ordered before specimen is collected</td>
</tr>
<tr>
<td>DRUGALLERGY</td>
<td>Patient allergic to ordered medication</td>
</tr>
<tr>
<td><strong>DRUGDRUG</strong></td>
<td>Ordered medication interacts with current medication</td>
</tr>
<tr>
<td><strong>DRUGDUP</strong></td>
<td>Ordered medication is a duplicate of current medication</td>
</tr>
<tr>
<td><strong>DRUGFOOD</strong></td>
<td>Ordered medication interacts with patient’s diet</td>
</tr>
<tr>
<td><strong>LH_ADVSR_VTE_2</strong></td>
<td>Prompts for a venous thromboembolism screening</td>
</tr>
<tr>
<td><strong>UAB_ADE16_NITRO_ACIDOSIS</strong></td>
<td>Nitroprusside medication ordered on a patient with contraindicated labs</td>
</tr>
<tr>
<td><strong>UAB_ADE1_HT_WT_ALLERGY</strong></td>
<td>Requires documentation of a patient’s allergies, height, and weight</td>
</tr>
<tr>
<td><strong>UAB_ADE2_RCMDCMKIDNEY</strong></td>
<td>Radiology study with contrast ordered on a patient with contraindicated labs</td>
</tr>
<tr>
<td><strong>UAB_ADE6_DIGOXINLAB</strong></td>
<td>Digoxin-type medication ordered on a patient with contraindicated labs</td>
</tr>
<tr>
<td><strong>UAB_ADE8_METFORM_ACIDOSIS</strong></td>
<td>Metformin prescribed to a patient with contraindicated labs in the last 14 days</td>
</tr>
<tr>
<td><strong>UAB_ADE9_ANTIHISTEDL_V2</strong></td>
<td>Anticholinergic or antihistamine medication ordered on a patient over 65</td>
</tr>
<tr>
<td><strong>UAB_RULE101_MEDS_10_DAYS</strong></td>
<td>Medication order signed with start date more than 10 days in the past</td>
</tr>
<tr>
<td><strong>UAB_RULE101_MED_10DAYS_MM</strong></td>
<td>Medication added to scratch pad with start date more than 10 days in the past</td>
</tr>
<tr>
<td><strong>UAB_RULE105_DUP_PSYCH_MED</strong></td>
<td>Two or more psychotropic agents ordered on scratch pad</td>
</tr>
<tr>
<td><strong>UAB_RULE111_SULF UREA_ALT</strong></td>
<td>Sulfonlurea ordered on a patient &gt;50 years old with creatinine clearance &lt; 60</td>
</tr>
<tr>
<td><strong>UAB_RULE158_ED_FLOOR_ADM</strong></td>
<td>Suggests admission to ICU for patients matching lab values or vital signs</td>
</tr>
<tr>
<td><strong>UAB_RULE175_TRANSFUSION</strong></td>
<td>Transfusion ordered on a patient with a hemoglobin ≥ 7 g/dL</td>
</tr>
<tr>
<td><strong>UAB_RULE184_CHESTAPALERT</strong></td>
<td>Chest x-ray ordered with existing one less than 3 days old</td>
</tr>
<tr>
<td><strong>UAB_RULE186_PLATELET_TRAN</strong></td>
<td>Platelet order placed without results within 30 days or without indication</td>
</tr>
<tr>
<td><strong>UAB_RULE209_METHOTREX_CRC</strong></td>
<td>Methotrexate ordered on a patient with creatinine clearance ≤ to 50</td>
</tr>
<tr>
<td><strong>UAB_RULE22_RAD_HVC_WT_LMT</strong></td>
<td>Radiology study is ordered on a patient weighing &gt; 225 kg</td>
</tr>
<tr>
<td><strong>UAB_RULE22_RAD_NM_WT_LMT</strong></td>
<td>Nuclear medicine study on a patient weighing over 158.76 kg</td>
</tr>
<tr>
<td><strong>UAB_RULE22_RAD_PET_WT_LMT</strong></td>
<td>PET scan ordered on a patient over a pre-defined weight limit</td>
</tr>
<tr>
<td><strong>UAB_RULE22_RAD_WT_HIGHLAN</strong></td>
<td>Radiology study ordered on a patient over equipment weight limit</td>
</tr>
<tr>
<td><strong>UAB_RULE22_UABH_C_CT_WTLM</strong></td>
<td>Radiology study ordered on a patient over equipment weight limit</td>
</tr>
<tr>
<td><strong>UAB_RULE23_MEPERIDINECRCL</strong></td>
<td>Meperidine ordered on a patient with a previous creatinine clearance or glomerular filtration rate &lt; 50 mL/min.</td>
</tr>
<tr>
<td><strong>UAB_RULE232_CDIFF_TOXIN</strong></td>
<td>C diff toxin test ordered on a patient taking a laxative</td>
</tr>
<tr>
<td><strong>UAB_RULE255_MYASTHENIAGRA</strong></td>
<td>Contraindicated medication ordered on a patient with myasthenia gravis</td>
</tr>
<tr>
<td><strong>UAB_RULE257_NAC_TYL</strong></td>
<td>Tylenol ordered on a patient taking acetylcysteine</td>
</tr>
<tr>
<td><strong>UAB_RULE268_HALOPERDL65YO</strong></td>
<td>Haloperidol is prescribed to a patient &gt; 65 years old</td>
</tr>
<tr>
<td><strong>UAB_RULE27_HIT_AB</strong></td>
<td>Heparin ordered on a patient with anti-heparin antibodies</td>
</tr>
<tr>
<td><strong>UAB_RULE282_RAD_CTLUNG_SC</strong></td>
<td>Requires a completed form before performing a lung cancer CT screen</td>
</tr>
<tr>
<td><strong>UAB_RULE28_STERILE_WATER</strong></td>
<td>Sterile water without an electrolyte additive ordered</td>
</tr>
<tr>
<td><strong>UAB_RULE35_WAR_INF_SIGN</strong></td>
<td>Warfarin ordered and signed without an INR within 7 days</td>
</tr>
<tr>
<td><strong>UAB_RULE35_WAR_INR_SCRATC</strong></td>
<td>Warfarin ordered on scratchpad without an INR within 7 days</td>
</tr>
<tr>
<td><strong>UAB_RULE35_WAR_INR_SR_AMB</strong></td>
<td>Warfarin ordered and signed without an INR within 7 days (ambulatory rule)</td>
</tr>
<tr>
<td><strong>UAB_RULE37_PO_CONV_ALERT</strong></td>
<td>Recommends IV to oral medication conversion if a patient’s diet order suggests the action would be appropriate.</td>
</tr>
<tr>
<td><strong>UAB_RULE61_MRI_SAFETY_SCR</strong></td>
<td>Requires completed safety form before signing MRI order</td>
</tr>
<tr>
<td><strong>UAB_RULE62_MEPERIDINECRCL</strong></td>
<td>Meperidine ordered on a patient with a previous creatinine clearance or glomerular filtration rate &lt; 50 mL/min.</td>
</tr>
<tr>
<td><strong>UAB_RULE63_US_GUIDED_LAB</strong></td>
<td>Ultrasound procedure ordered without platelet lab results</td>
</tr>
<tr>
<td><strong>UAB_RULE63_US_GUIDED_GB_C</strong></td>
<td>Ultrasound procedure ordered without platelet lab results</td>
</tr>
<tr>
<td><strong>UAB_RULE90_DOPAMINERGIC</strong></td>
<td>Contraindicated medication ordered for a patient taking dopamine</td>
</tr>
<tr>
<td><strong>UAB_RULE90_DOPAMINERGIC_2</strong></td>
<td>Contraindicated medication ordered for a patient taking dopamine</td>
</tr>
<tr>
<td><strong>UAB_RULE90_DOPAMIN_MEDM</strong></td>
<td>Contraindicated medication ordered for a patient taking dopamine</td>
</tr>
<tr>
<td><strong>UAB_RULE90_DOPAMIN_MEDM2</strong></td>
<td>Contraindicated medication ordered for a patient taking dopamine</td>
</tr>
<tr>
<td><strong>UAB_RULE92_ACE_PHER_ALERT</strong></td>
<td>ACE inhibitor is ordered for a patient scheduled for plasmapheresis</td>
</tr>
<tr>
<td><strong>UAB_RULE92_ACE_PHER_MEDM</strong></td>
<td>ACE inhibitor is ordered for a patient scheduled for plasmapheresis</td>
</tr>
</tbody>
</table>
Patient Data
In addition to acquiring all alerts for the seven-year period, all patient data recorded during that time period were retrieved. The information included the date of birth for age calculation along with basic demographic information for the analysis such as race and sex. Duplicate patients (based on possession of identical medical record numbers) in the data set were combined while retaining the most specific demographic information. For example, if two duplicate patient records had a race of “Unknown” and “African American”, the final aggregated patient record’s race was set to “African American”.

Early Warning Score
After combining the appropriate patient information with each alert event, the early warning score was added. Since the patient’s vitals, and by extension the EWS, are not guaranteed to be measured at the time an alert event triggers, a reasonable estimate was acquired using the following procedure. For each alert event, a time window consisting of 12 hours before and 2 hours after was created. Then, the EWS within that window and occurring closest to the time the alert triggered was taken to be the patient’s EWS at the time of the event.

Using 12 hours or less prior to the alert increased the likelihood that the EWS would be from the same hospital stay as the alert event and hence more likely to accurately reflect the patient’s current status. Likewise, searching two hours after the alert made it possible to detect EWSs occurring slightly after the alert that would more accurately reflect the patient’s status compared to an EWS taken ten hours prior to the alert. Any alert events without an EWS within the time window were excluded from analysis.

Statistical Analysis
In order to determine if a patient’s overall health status as indicated by the EWS was predictive of an alert’s override likelihood, a logistic regression model with general estimating equations (GEE) was employed. Logistic regression allows modeling of the binary outcome: override or non-override. While capable of modeling this type of outcome, logistic regression expects independence of each alert event. However, an alert can occur on the same patient, breaking the independence assumption of the model. The addition of GEE to the analysis corrected for this non-independence by estimating correlation between alert events.

Following de-identification of the data, the logistic regression model with GEE was performed using SAS 9.4. Each alert was tested in a separate GEE model in order to make interactions between a specific alert and the EWS levels more easily interpretable (i.e. allow negative versus positive correlations between EWS and an alert to be more recognizable). Patient demographics served as covariates in the model in order to account for any variation that might occur due to these categories. Additionally, the amount of time between the EWS measurement and the alert event might affect the accuracy of EWS and, as a result, the override likelihood. Therefore, both the time between the EWS and the alert along with a time / EWS interaction term were included. A conservative Type III analysis was used for each categorical variable. Finally, due to the multiple models being run simultaneously, a Bonferroni correction was applied to all p-values. As 46 models were run, an alpha value of 0.0011 (0.05 / 46) was considered to be significant for Type III analyses.

Results
The UAB EHR contains 52 alerts that offer physicians an override option (see Table 2). Over the seven-year period (starting at the launch of UAB’s EHR), these alert rules represent a total of 4,043,431 unique alert events with approximately 17.37% overridden on average. The events per alert rule range from 1 to 1,998,811 (see Figure 1). Even after excluding rules that occurred less than 1000 times, override percentages ranged from 0.02% or 100% (see Figure 1).

Additionally, the basic demographic information of the UAB patient population was retrieved for inclusion in the models. This information included patient sex, race, and age at the time of the alert. The breakdown of the patient population during the time that alert event data was pulled is shown in Table 3.

The appropriate patient demographic information was combined with each alert event. Then, each alert event was combined with a patient’s EWS closest to the event and occurring within 12 hrs. before the event or 2 hrs. after the
event (see Methods for details). Alert events without an EWS within this time window were excluded leaving a total of 46 alert rules spanning 443,019 events from the original 52 alert rules and 4,043,431 events.

These 46 alerts were then individually analyzed with logistic regression incorporating GEE. Due to limitations of the data for some alerts, 20 were unable to be analyzed for various reasons. The analysis removed eight alerts due to one of the variables having only one value (e.g. all alert events in the data set were overridden), and it removed twelve alerts for insufficient variation in one or more variables. The alerts events in the data with only one type of override status had either a low number of events in the combined data set, extreme override frequencies (high or low), or both.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Category</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female</td>
<td>921,874 (54.84%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>754,637 (44.89%)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>4,384 (0.26%)</td>
</tr>
<tr>
<td>Race</td>
<td>White</td>
<td>109,209 (6.5%)</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>54,317 (3.23%)</td>
</tr>
<tr>
<td>Age</td>
<td>Mean (yr.)</td>
<td>54.7</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation (yr.)</td>
<td>17.1</td>
</tr>
</tbody>
</table>

The results of the remaining alerts are shown in Table 4. After using a Bonferroni correction ($\alpha = 0.05/46 = 0.0011$) to account for using multiple models (one per alert), EWS only predicts override frequency for three alerts: DRUGDRUG, DRUGDUP, and UAB_RULE184_CHESTAPALERT. For two of these alerts (DRUGDRUG and UAB_RULE184_CHESTAPALERT), the time term is also significant in addition to the separate EWS and time terms, suggesting that the time between the EWS measurement and alert event affects the utility of EWS in predicting an alert’s override. Additionally, the interaction term between time and EWS reached significance for a few of the alerts, including the DRUGDRUG and the UAB_RULE184_CHESTAPALERT. This information suggests that both the time between the alert event and the EWS measurement and the EWS itself combine to produce different likelihoods of an alert overrides.

These results can further be visualized in Figure 2. This representation depicts the override percentage of an alert at each EWS level. Although there is not a drastic increase or decrease in override percentage with increasing EWS for the three significant alerts, the amount of data analyzed allows detection of smaller effects. Additionally, as in the UAB_RULE184_CHESTAPALERT alert, the baseline override percentage is already high, and does slightly increase with increasing EWS.

Discussion

EHRs use alerts to warn providers of potential errors or provide suggestions on patient care. The large number of alert events that occur have potentially decreased their utility in certain instances and introduced other issues including alert fatigue. Although most EHRs give the provider a default list of override reasons, these do not always accurately reflect the reasoning behind the override decision. Additionally, appropriate overrides are typically the result of the clinician considering an aspect of the clinical context not considered by the alert logic. This study investigated the potential of using a patient’s health status (EWS) to predict an alert’s override. Presumably, clinicians might override minor workflow suggestions as a patient’s health status deteriorates and adhere to severe alerts (e.g. drug-renal alerts) more frequently as a patient’s EWS increases. Although overrides and false positive alerts are not identical, many alerts are appropriately overridden. Therefore, targeting override prediction is a good first step to removing potentially unnecessary alerts. In order to be as comprehensive as possible, all overridable alerts at UAB were analyzed; however, a patient’s health status might not affect a clinician’s decision for all alerts. For example, while the UAB_ADE8_METFORM_ACIDOSIS alert warning of a contraindicated prescription of metformin might cause a clinician to consider the severity of the patient’s condition, the
UAB_ADE1 HT WT_ALLERGY alert requiring patient allergy, height, and weight information be added to the record might not have the same effect.

Although EWS did not achieve significance for explaining overrides for many of the alerts, it did for three: drug-drug interaction (DRUGDRUG), drug duplicate orders (DRUGDUP), and a chest X-ray advisory (UAB_RULE184_CHESTAPALERT). Regarding the two drug alerts, a clinician might consider the medication more important for the deteriorating patient than the potential of an interaction. Indeed, a previous study showed that the two most common override reasons for drug-drug interaction alerts were ‘clinically irrelevant alert’ and ‘benefit assessed to be greater than the risk’, and this result is reflected in the UAB EHR as the most common override reasons for both the DRUGDRUG and DRUGDUP alert are ‘interaction noted, will take precautions’ and ‘essential therapy, will take precautions’. Similarly, the UAB_RULE184_CHESTAPALERT alert warns clinicians that a chest x-ray has already been performed on the patient within 3 days and suggests not ordering another one. A deteriorating patient would likely increase the likelihood of overriding such an alert as shown in the model results (see Figure 2).

It should be noted that the increase override percentage shown Figure 2 for these alerts is small and might not offer a threshold for EWS at which the alert completely loses utility. However, this result is likely due to the already high override percentage.

<table>
<thead>
<tr>
<th>Alert Name</th>
<th>EWS p-value from Type 3 GEE</th>
<th>Time p-value from Type 3 GEE</th>
<th>EWS * Time p-value from Type 3 GEE</th>
<th>EWS Coefficient from GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLERGYDRUG</td>
<td>0.12</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>-0.01</td>
</tr>
<tr>
<td>DRUGALLERGY</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>DRUGDRUG</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.04</td>
</tr>
<tr>
<td>DRUGDUP</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>DRUGFOOD</td>
<td>0.04</td>
<td>0.67</td>
<td>0.67</td>
<td>0.17</td>
</tr>
<tr>
<td>UAB_ADE16_NITRO_ACIDOSIS</td>
<td>0.02</td>
<td>0.18</td>
<td>0.18</td>
<td>0.24</td>
</tr>
<tr>
<td>UAB_ADE2_RCMDMKIDNEY</td>
<td>0.2</td>
<td>0.98</td>
<td>0.98</td>
<td>-0.03</td>
</tr>
<tr>
<td>UAB_ADE6_DIGOXINLAB</td>
<td>0.02</td>
<td>0.99</td>
<td>0.99</td>
<td>0.08</td>
</tr>
<tr>
<td>UAB_ADE8_METFORM_ACIDOSIS</td>
<td>0.98</td>
<td>0.75</td>
<td>0.75</td>
<td>0</td>
</tr>
<tr>
<td>UAB_ADE9_ANTIHISTEDL_V2</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>UAB_RULE101_MEDS_10_DAYS</td>
<td>0.52</td>
<td>0.82</td>
<td>0.82</td>
<td>-0.05</td>
</tr>
<tr>
<td>UAB_RULE111_SULF_UREA_ALT</td>
<td>0.17</td>
<td>0.51</td>
<td>0.51</td>
<td>0.24</td>
</tr>
<tr>
<td>UAB_RULE158_ED_FLOOR_ADM</td>
<td>0.07</td>
<td>0.67</td>
<td>0.67</td>
<td>-0.06</td>
</tr>
<tr>
<td>UAB_RULE175_TRANSFUSION</td>
<td>0.54</td>
<td>0.06</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>UAB_RULE184_CHESTAPALERT</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.17</td>
</tr>
<tr>
<td>UAB_RULE186_PLATELET_TRAN</td>
<td>0.99</td>
<td>0.04</td>
<td>0.04</td>
<td>0</td>
</tr>
<tr>
<td>UAB_RULE22_RAD_NM_WT_LMT</td>
<td>0.45</td>
<td>0.54</td>
<td>0.54</td>
<td>-0.51</td>
</tr>
<tr>
<td>UAB_RULE268_HALOPERDL65YO</td>
<td>0.03</td>
<td>0.3</td>
<td>0.3</td>
<td>0.24</td>
</tr>
<tr>
<td>UAB_RULE35_WAR_INF_SIGN</td>
<td>0.36</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>UAB_RULE35_WAR_INR_SCRATC</td>
<td>0.41</td>
<td>0.63</td>
<td>0.63</td>
<td>-0.06</td>
</tr>
<tr>
<td>UAB_RULE37_PO_CONV_ALERT</td>
<td>0.01</td>
<td>0.24</td>
<td>0.24</td>
<td>-0.03</td>
</tr>
<tr>
<td>UAB_RULE61_MRI_SAFETY_SCR</td>
<td>0.37</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>UAB_RULE62_MEPERIDINECRCL</td>
<td>0.16</td>
<td>0.19</td>
<td>0.19</td>
<td>-0.1</td>
</tr>
<tr>
<td>UAB_RULE90_DOPAMINERGIC</td>
<td>0.37</td>
<td>0.05</td>
<td>0.05</td>
<td>0.09</td>
</tr>
<tr>
<td>UAB_RULE90_DOPAMINERGIC_2</td>
<td>0.51</td>
<td>0.63</td>
<td>0.63</td>
<td>0.11</td>
</tr>
<tr>
<td>UAB_RULE92_ACE_PHER_ALERT</td>
<td>0.83</td>
<td>0.55</td>
<td>0.55</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 4. Logistic Regression with GEE results for the EWS predictor and the EWS * Time interaction term.
Figure 1. The distribution of alert events in the UAB EHR broken down by override status. Alert rules were limited to those that provided clinicians an override option. The length of the bars indicates the log transformed number of the total alert events triggered by the alert rule on the y-axis. The two colors of the bar reflect the ratio of overridden to non-overridden alert events (alerts events that clinicians chose the alternative option to overriding) with the percentage of each alert rule’s overrides displayed to the right of the bar. Many of these alerts have override rates above 70% showing that the UAB institution has similar override rates to the 49 – 96% reported by other institutions.
Figure 2. The effect of EWS on the override frequency of the 26 alerts analyzed by logistic regression with GEE. Override frequency per alert per EWS level was calculated and plotted with an estimated line using simple linear regression. Each data point for a specific EWS and override frequency was duplicated to reflect the number of alert events and weight the information. Thus, the trend lines indicate whether an alert is more or less likely to be overridden with deteriorating patient status. The asterisks indicate alerts in which EWS is a significant predictor of alert overrides. Other alerts have some increase or decrease in override frequency with an increasing EWS suggesting that patient status might play a small part in their override likelihood.
Additionally, the interaction term between time and EWS reached significance for a few of the alerts. This result might be an indication of the accuracy of EWS at the time of the alert. Although the method of selecting the EWS was intended to choose the measurement most reflective of the patient’s status at the time of the alert event, this result might not have been achieved as closely as desired. Thus, the time between the alert event and the EWS measurement might cause EWS to differentially predict alert overrides.

However, EWS did not significantly correlate with the likelihood of an override for many of the alerts. Some of the reasons for this might be due to the fact that a patient’s health status simply does not factor into a physician’s decision to override the alert as discussed above. However, another reason might be that many of the alerts already have an override rate above 75% (see Figure 1) limiting the ability of the models to detect an effect. Additionally, it should be noted that medical alerts widely vary in the decision support they provide. Some simply suggest cost-saving measures as in the chest x-ray alert (UAB_RULE184_CHESTAPALERT). Others provide minor suggestions for patient care such as warning of false positives on a lab test (UAB_RULE232_CDIFF_TOXIN). Some provide critical care advice like the alert warning of a methotrexate prescription on a patient under a specified creatinine clearance (UAB_RULE209_METHOTREX_CRC). Given this variation, a single patient characteristic is unlikely to provide a strong predictor for the overrides of all alerts.

As some of the alerts did approach significance for EWS as a predictor, more covariates or additional predictors are needed to account for the variance in the models. For example, the UAB_ADE9_ANTIHISTELD_V, which warns about prescribing an anticholinergic or antihistamine medication to a patient over 65, might benefit from accounting for previously prescribed and tolerated medications. For other alerts, such as the UAB_RULE184_CHESTAPALERT, previous diagnoses or procedures might be effective at determining whether a clinician would override the alert. Newly recorded procedures might prompt a physician to order a patient X-ray in order to track the individual’s current status.

Alternatively, given there is a need to remove clinically unnecessary alerts, another more effective approach might be to implement a machine learning model to predict overrides. This methodology would incorporate multiple EHR variables along with complex interactions and might come closer to solving the immediate need of removing unnecessary alert events.

Conclusion

High rates of false positive alerts continue to plague clinicians despite decades of work to reduce them. We have explored a methodology for predicting such alerts based on patient data in the EHR available just prior to the time of the alert. The use of the EWS for this purpose may be one small step toward alert reduction. The EHR is replete with many other possible targets for the methodology we describe here, offering the potential for a giant leap forward in addressing this decades-old problem.

Acknowledgements

Drs. Kennel and Cimino are supported by research funds from the UASOM Informatics Institute. Dr. Cimino is also supported in part by research funds from the Center for Clinical and Translational Sciences (CCTS) under grant UL1TR001417 from the National Center for the Advancement of Translational Science (NCATS).

References


ABSTRACT
Sleep change is commonly reported in Alzheimer’s disease (AD) patients and their brain wave studies show decrease in dreaming and non-dreaming stages. Although sleep disturbance is generally considered as a consequence of AD, it might also be a risk factor of AD as new biological evidence shows. Leveraging the National Sleep Research Resource (NSRR), we built a unique cohort of 83 cases and 331 controls with clinical variables and electroencephalography (EEG) signals. Supervised tensor factorization method was applied for this temporal dataset to extract discriminative sleep patterns. Among the 30 patterns extracted, we identified 5 significant patterns (4 patterns for AD likely and 1 pattern for normal ones) and their visual patterns provide interesting linkage to sleep with repeated wakefulness, abnormal REM sleep, and insomnia. This study is preliminary but findings are interesting, which is a first step to provide quantifiable evidences to measure sleep as a risk factor of AD.

INTRODUCTION
Alzheimer’s disease (AD) has become a major public health concern because of its increasing prevalence, chronicity, caregiver burden, and high personal and financial costs of care. About 25% to 40% of AD patients suffer from disturbances of sleep, such as insomnia at night, rapid eye movement (REM) sleep behavior disorder, agitated behavior at sunset and excessive sleeping during the daytime [1]. In addition, sleep apnea syndrome is prevalent in AD patients [2], which is associated with APOE4 that is a well-known risk factor of AD [3]. Those sleep problems occur early on in the course of AD, consistent with the finding that brain regions involved in sleep are affected during the development of AD. This fact implies that these sleep problems can be a preclinical marker for the development of AD [4]. A quantitative tool to measure such a sleep disturbance is electroencephalography (EEG), which is one of the most popular brain signals that records brain’s spontaneous electrical activity on different regions of scalp over a period of time. This signal usually transforms to multiple waves according to frequency domain (e.g., Alpha, Theta, Delta). EEG can quantify sleep patterns. In the normal young adult, sleep consists of five cycle stages:

I. Stage 1 (5%): a period of transition from wakefulness (Alpha wave, 8-10.5 Hz) to light sleep (Theta wave, 4 - 8 Hz);
II. Stage 2 (50%): light sleep with Theta wave (4 - 8 Hz);
III. Stage 3 and Stage 4 (20%): slow-wave activity (SWA) or deep sleep with Delta wave (1 - 4 Hz);
IV. Stage 5 (25%): REM sleep with a fast desynchronized EEG containing Alpha (8 - 10.5 Hz), Beta (15 - 30 Hz), and Theta (4 - 8 Hz) waves [5].

With increasing age, sleep becomes “lighter” in that the percentage of Stage 1 increases and the percentage of Stage 3 and 4 decreases. Patients with AD show an increased number and duration of awakenings [6,7], consequently EEG recordings show that percentage of Stage 1 increases and percentage of Stage 2 and SWA (Stage 3 and 4) decreases more than normal aging does. Another interesting EEG pattern of AD is that the amount of REM sleep decreases and this change is mostly seen in later stages of AD [7]. Also, EEG slowing is observed prominently in REM sleep (other than in the awake EEG) [8,9]. These findings are driven by researchers who have been visually compared EEG between normal elderly and AD patients with their human naked eyes [9]. Unfortunately, this ad-hoc approach is not scalable to large samples and have difficulty in exploring a large combination of patterns associated with temporal changes on frequency bands and brain regions across sleep cycles [6]. Lack of quantifiable measurement can be another issue, which might lead to inconsistency in judgement. Therefore, a computational method that directly learns sleep patterns from observational data becomes more important in discovering meaningful and generalizable latent patterns of sleep EEG.

Dimensionality reduction is a widely-used method to discover such underlying latent space from high dimensional data. Particularly, linear models, such as principal component analysis, linear discriminant analysis, and nonnegative matrix factorization, represent the observed data as a weighted linear sum of latent dimensions and have been widely
applied in EEG analysis [10]. Nonnegative tensor factorization (NTF) is another powerful linear model, which decomposes high dimensional data (such as time-frequency representation of EEG from multiple samples) into linear sum of basis components (which are interpreted as patterns). NTF has been applied to several medical domains, such as phenotyping [11,12] from electronic medical records, temporal patterns from behavioral log [13], and event-related EEG [14–16]. To derive discriminative phenotypes or patterns with respect to certain clinical outcome of interest, an extension of NTF, so called supervised NTF, has been proposed and enforces the basis components to be discriminative to the clinical outcome of interest [12].

In this work, we aim to discover temporal patterns from sleep EEG to differentiate AD from the control (non-AD), using supervised NTF. Previous studies on discovering AD patients’ EEG patterns use only event-related EEG (not focused on sleep EEG, which capture very different characteristic of brain functionality). Also, they are unsupervised approach within AD populations (not supervised approach discriminating AD case and non-demented control) [17,18]. We focus on the sleep EEG and integrate the supervision term explicitly in our objective function to derive discriminative temporal patterns as computational phenotypes.

METHODS

We developed a computational framework to discover sleep EEG patterns that would be discriminative to AD. This framework consists of three steps: i) patient matching; ii) EEG transformation; and iii) tensor factorization (Figure 1).

**Figure 1.** Overview of computational framework to discover discriminative patterns from Sleep EEG. This framework consists of three steps: i) patient matching; ii) EEG transformation; and iii) tensor factorization. For each patient, the raw EEG signal was transformed into frequency-time representation using Fourier transform. The transformed EEG was in a matrix format. We stacked all the individual matrices from each patient into a 3-order tensor.

Datasets

We leveraged two datasets in NSRR [19][20], which contains large collections of de-identified clinical data elements and electrophysiological signals from over 10 NIH-funded sleep cohort studies. Clinical data elements include demographic information (e.g., age, gender, race), anthropometric parameters (e.g., body mass index, height), physiologic measurements (e.g., diastolic/systolic blood pressure, heart rate), medical history (e.g., atrial fibrillation, cancer, depression), medications (e.g., antidepressant, acetylcholinesterase inhibitors for AD) and sleep symptoms (e.g., excessive daytime sleepiness). Electrophysiological signals in the European Data Format (EDF) are the polysomnography recordings (overnight sleep), with various channels such as electroencephalogram (EEG) and electrocardiogram (ECG). The two datasets we used were the Sleep Heart Health Study (SHHS) [21] and MrOS Sleep Study (MrOS) [22][23][24], where SHHS contains a total of 5,804 adults aged 40 or older, and MrOS contains a total of 2,911 men aged 65 years or older.
Since these two datasets were not specifically collected to study the relationship between sleep and AD, we performed retrospective analyses as a secondary use of these two datasets. To determine whether each patient has onset of AD we utilized a clinical variable on a specific medication, acetylcholinesterase inhibitors. Acetylcholinesterase inhibitors is widely used drug for AD. SHHS has a clinical variable asking if the patient has taken acetylcholinesterase inhibitors for AD within two weeks of the study visit. MrOS has three clinical data elements related to AD or AD medication use: (1) Has a doctor or other health care provider ever told you that you had dementia or AD? (2) Are you currently being treated for dementia or AD by a doctor? and (3) AD medication use.

Cohort Selection using Propensity Score Matching
Among 1,959 (MrOS) and 1,893 (SHHS) patients with age > 65, we selected 331 controls normal (CN, non-AD) that match with 83 case (AD) with around 4:1 ratio in terms of potential confounding risk factors to AD. Due to complex nature of AD’s neurodegeneration, many factors contribute to the disease. In order to focus on the relationship of AD and EEG signals, we need to reduce effects of other confounding variables that affect the incidence of AD, such as demographic background (e.g., age, gender, and race) and potential risk factors (e.g., depression, hypertension, diabetes, cardiovascular disease, stroke). We utilized a statistical matching model called propensity score matching (PSM) [25], which matches estimators (i.e., confounders such as depression and hypertension) so that the distribution of EEG signals is independent of the outcome (i.e., AD), conditioned on the confounders [26]. The PSM algorithm finds the matched controls to the cases using two steps: i) computing propensity scores and ii) finding similar controls to each case using propensity scores. First, we obtained propensity scores as probability of AD computed from logistic regression, in which AD incidence is the binary label and confounders are predictors. The propensity scores (PSM) algorithm finds the best combination of n and r that leads to the biggest sample size within a tolerable bias.

EEG Frequency Bands Representation
We derived time-frequency representation from the raw EEG signals. EEG signals have various behaviors in different frequency bands, and the characteristics of different frequency bands have been reported in previous work [27–29]. For example, previous studies demonstrated that spectral power is an important feature for sleep stage research [30], such as: the power of lower frequencies of EEG becomes stronger with the increasing depth of sleep; the EEG signal of Stage 1 has more power between 2-7 Hz; Stage 2 can be characterized by the presence of sleep spindles band (12-15 Hz), and SWA is defined when there are low frequency (less than 2 Hz) waves [31].

We first divided the entire sleep EEG records into non-overlap 30-second epochs during 8 hours (i.e., 8 hours / 30 seconds = 960 epochs). EEG signals after the 8 hours were truncated. After separating EEG time dimension into 960 epochs, we separated the EEG values into several waves depending on its frequency bandwidths using power spectral analysis. The most standard frequency bands include Delta wave, Theta wave, Alpha wave (Table 1). Fast Fourier transform (FFT) is a direct and commonly used spectral estimation method for the EEG frequency analysis [28]. The discrete Fourier transform (DFT) is an efficient numerical algorithm to perform Fourier transform in many practical applications. Let us denote N is the number of EEG values in one epoch and $x_n$ as a single EEG value in one epoch. The DFT of a sequence of N values $\{x_n\} := x_0, x_1, \ldots, x_{N-1}$ is defined as:

$$X_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n \cdot e^{-i(2\pi nk/N)}$$

where $\{X_k\} = X_0, X_1, \ldots, X_{N-1}$ is a transformed sequence with complex numbers of $\{x_n\}$ for $k = 0, \ldots, N - 1$, and $e$ is the Euler’s number. The power spectrum is then obtained with

$$P_k = \|X_k\|^2.$$

The $P_k$ values for $k = 0, \ldots, N - 1$ are then separated according to the 12 frequency bands (Table 1). For example, $P_k$ values lying in every $\frac{1}{3}$ to $\frac{1}{2}$ seconds (4 - 8 Hz) correspond to Theta wave. For each frequency band, we summed all the $P_k$ values and used it as an amplitude value. So, in total 12 frequency amplitude values were generated per every epoch.
Table 1. Extracted EEG frequency bands.

<table>
<thead>
<tr>
<th>Name</th>
<th>Hz</th>
<th>Name</th>
<th>Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow-Oscillations (SO)</td>
<td>[0.5, 1.5]</td>
<td>Sigma</td>
<td>[12.0, 15.0]</td>
</tr>
<tr>
<td>Slow-wave activity (SWA)</td>
<td>[0.5, 5.5]</td>
<td>Slow Sigma</td>
<td>[12.0, 13.5]</td>
</tr>
<tr>
<td>Delta</td>
<td>[1.0, 4.0]</td>
<td>Fast Sigma</td>
<td>[13.5, 15.0]</td>
</tr>
<tr>
<td>Theta</td>
<td>[4.0, 8.0]</td>
<td>Beta1</td>
<td>[15.0, 20.0]</td>
</tr>
<tr>
<td>Alpha</td>
<td>[8.0, 10.5]</td>
<td>Beta2</td>
<td>[20.0, 30.0]</td>
</tr>
<tr>
<td>Spindle</td>
<td>[10.5, 14.5]</td>
<td>Gamma</td>
<td>[30.0, 60.0]</td>
</tr>
</tbody>
</table>

Discovering Biomarkers using Tensor Factorization

Using the EEG time-frequency representation of each patient, we constructed an observed tensor and extracted discriminative temporal patterns by applying supervised NTF to the observed tensor.

Construct Tensor. To derive patterns from time-frequency representation of EEG, we represented each patient’s signal amplitude as a matrix (12 frequency bands × 960 time slots), see Figure 2. We applied log2-transform to the amplitude to make the distribution follow normal distributions and facilitate stable tensor factorization. We stacked those matrices into a third-order tensor $O$ with a shape of (83 AD + 331 CN) patients × 12 frequency bands × 960 time slots (Tensor is a generalization of matrix. Order of a tensor is the number of dimensions. A first-order tensor is a vector, a second-order tensor is a matrix, and tensors of order three or higher are called high-order tensors).

Figure 2. EEG time-frequency representation of all patients forms a tensor. AD= Alzheimer’s disease, CN= Control.

Regularized Nonnegative Tensor Factorization. The most widely used tensor decomposition is the CP method [32]. A third-order tensor $O$ with shape of $I \times J \times K$ is rank-one if it is an outer product of three vectors $a, b, c$, i.e., $O = a \circ b \circ c$ where $\circ$ means the vector outer product. $O_{ijk}$, the element at $(i, j, k)$ in the tensor $O$, is computed as product of elements in the vector, i.e., $O_{ijk} = a_i b_j c_k$. Tensor factorization (TF) is a dimensionality reduction approach that represents the original tensor as latent dimensions. The CP model approximates the original observed tensor $O$ as a linear combination of rank-one tensors [32]; that is, a third-order tensor $O$ is decomposed as minimizing difference between observed tensor and approximated tensor as

$$L = ||O - \sum_{r=1}^{R} a_r \circ b_r \circ c_r||^2$$

where a positive integer $R$ is the rank, $a_r, b_r, c_r$ are $r$-th column vectors in matrix $A, B, C$ with shape of $I \times R, J \times R, K \times R$, respectively. Here, $A, B, C$ are called as factor matrices. When tensor $O$ contains non-negative data (such as amplitude, intensity or counts), we set non-negative constraints $A, B, C \geq 0$ for interpretability of latent dimensions, which is so called nonnegative tensor factorization (NTF). Since our objective is to derive
discriminative patterns, we used supervised version of TF, which adds a supervised regularizer that encourages the 
patterns to be separated according to AD vs CN [12]. The supervised TF adds logistic regression regularizer as

$$L - \mu \cdot \log P(A, y|\theta) = L - \mu \cdot \frac{1}{1 + \exp(-y \cdot \theta A)}$$

where $\mu$ is a weight parameter to balance the tensor error and the loss on regularizer, $\theta$ is parameter for logistic 
regression, and $y$ is label ($y = 1$ if AD; -1 if CN). To further enhance the interpretability via compact patterns, we 
also added $l_1$-norm regularizer to the factor matrix. The $l_1$ regularizer shrinks the less important coefficients to zero, 
as improving interpretability of model:

$$L + \lambda \cdot (||A||_1 + ||B||_1 + ||C||_1)$$

where $\lambda$ is a weight parameter to balance the tensor error and the $l_1$-norm loss. Therefore, our final objective function 
is adding both of regularizers:

$$L - \mu \cdot \log P(A, y|\theta) + \lambda \cdot (||A||_1 + ||B||_1 + ||C||_1).$$

**Discover Patterns using NTF.** We applied this regularized NTF to our EEG dataset. We let factor matrix $A$, $B$, and 
$C$ represent patients (AD + CN), frequency bands, and time slots, respectively. The $R$ latent dimensions in factor 
matrices $A$, $B$, and $C$ represent temporal patterns of EEG. Each EEG pattern consists of a set of frequency bands and 
a set of time slots. Individual frequency band or time slot contributes to the pattern with different extent of 
membership, and the amount of contribution is stored in columns of $B$ and $C$. Likewise, individual patients have 
characteristics of the $R$ patterns with different extent of membership. That is, the column vector of $B$ defines how 
much the frequency bands are involved in each pattern, and the column vector of $C$ defines how much the time slots 
are involved in each pattern. The row vector of $A$ defines how much patients participate (or have membership) in the 
characteristic of each pattern.

**EXPERIMENTS AND RESULTS**

**Propensity Score Matching**

To create our training cohort for regularized NTF, we obtained controls for the cases from our data sources (MrOS, 
SHHS). We used propensity score matching to reduce the effect of age, race, common risk factors (hypertension, 
depression, stroke, diabetes) while trying to keep as much cases as possible. After matching, we lost a few cases in 
order to reduce the bias. We ended up with 83 cases and 331 controls at a roughly 1:4 ratio. Tables 2a and 2b 
summarize the statistics on features before and after the matching.

**Table 2a: Statistics of MrOS dataset before and after matching**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Original Dataset</th>
<th>PS Matched Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AD cohort</td>
<td>Control cohort</td>
</tr>
<tr>
<td>Number of Patients</td>
<td>86</td>
<td>2,397</td>
</tr>
<tr>
<td>Age</td>
<td>80.14±5.36</td>
<td>77.6±5.53</td>
</tr>
<tr>
<td>Race</td>
<td>86%</td>
<td>88.6%</td>
</tr>
<tr>
<td>Hypertension</td>
<td>50%</td>
<td>81.6%</td>
</tr>
<tr>
<td>Depression</td>
<td>5.8%</td>
<td>3.7%</td>
</tr>
<tr>
<td>Stroke</td>
<td>10.5%</td>
<td>6.3%</td>
</tr>
<tr>
<td>Diabetes</td>
<td>14.0%</td>
<td>18.7%</td>
</tr>
</tbody>
</table>
Table 2b: Statistics of SHHS dataset before and after matching

<table>
<thead>
<tr>
<th>Feature</th>
<th>Original Dataset</th>
<th>PS Matched Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AD cohort</td>
<td>Control cohort</td>
</tr>
<tr>
<td>Number of Patients</td>
<td>33</td>
<td>1,860</td>
</tr>
<tr>
<td>Age</td>
<td>79.6±6.9</td>
<td>73.5±7.83</td>
</tr>
<tr>
<td>Gender - Female</td>
<td>48.5%</td>
<td>53.5%</td>
</tr>
<tr>
<td>Race - Caucasian</td>
<td>87.9%</td>
<td>88.9%</td>
</tr>
<tr>
<td>Race - African American</td>
<td>9.1%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Cardiovascular Disease</td>
<td>39.4%</td>
<td>29.2%</td>
</tr>
<tr>
<td>Hypertension</td>
<td>51.5%</td>
<td>95.6%</td>
</tr>
<tr>
<td>Stroke</td>
<td>6.1%</td>
<td>4.5%</td>
</tr>
</tbody>
</table>

Discovering Biomarker using Tensor Factorization

We implemented the regularized NTF using Pytorch 11.4 with adaptive momentum estimation (ADAM) for optimization. We set the maximum number of iterations as 1,000. The running time was less than 15 secs with 3 parallel GPUs. We added dropout to logistic regression coefficients for robustness. After extensive parameter tuning, we set dropout rate=0.5, R=30, \( \mu = 0.1 \), and \( \lambda = 0.05 \).

Evaluating NTF methods. We computed discriminative power, sparsity, overlap of patterns that are derived from the NTF methods. We measured the discrimination by the area under the receiver operating characteristic curve (AUC) to classify AD and control. We measured compactness by sparsity and overlap of the temporal patterns. High sparsity means a few frequency bands or time period dominantly characterize each pattern whereas the other bands or time period are negligible, making interpretation of the patterns easy. The sparsity was computed as an averaged Gini index of involvement values in each pattern (i.e., the column vectors of \( B \) and \( C \)) [33]. The overlap measures the degree of overlapping between all pattern pairs [12]. Patterns with less overlap are more distinctly identified. The overlap is computed as an averaged cosine similarity between all pair of column vectors of \( B \) and \( C \). We also computed mean squared error (MSE) to evaluate how closely the derived patterns reflect the observed original data. We compared mean and standard deviation after ten repeated trials. We compared the discriminative power and compactness with different settings of regularizers:

- NTF: Basic NTF model without any regularizers (\( \mu = 0, \lambda = 0 \))
- NTF + logit: NTF with supervised regularizer based on logistic regression likelihood (\( \mu = 0.1, \lambda = 0 \))
- NTF +\( l_1\)norm: NTF with \( l_1\)norm (\( \mu = 0, \lambda = 0.05 \))
- NTF +\( l_1\)norm+logit: NTF with \( l_1\)norm and logistic regression likelihood (\( \mu = 0.1, \lambda = 0.05 \))

Table 3 summarizes interested measurements. We found that NTF with \( l_1\)norm and supervised term outperformed other baselines in terms of discriminative power and compactness. The NTF +\( l_1\)norm+logit showed the highest AUC, sparsity, and lowest overlap (Table 3). The \( l_1\)norm regularizer improved compactness (i.e., increased sparsity and decreased overlap). The supervised regularizer also improved discriminative power (i.e., increased AUC).

Evaluating Individual Patterns. We presented patterns that an NTF +\( l_1\)norm+logit model derives (with AUC=0.6981). After deriving \( R=30 \) patterns, we learned a logistic regression model using the \( R=30 \) patterns as predictors and selected patterns that separate AD and CN with statistical significance (Table 4: Coefficient). Five
Table 3. Discrimination and compactness comparison. We computed average and standard deviation after 10 repeated trials. The number of phenotypes $R = 30$.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AUC</th>
<th>Sparsity</th>
<th>Overlap</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTF ($\mu = 0$, $\lambda = 0$)</td>
<td>0.6716 (0.0321)</td>
<td>0.1799 (0.0078)</td>
<td>0.8936 (0.0094)</td>
<td>0.3061 (0.0038)</td>
</tr>
<tr>
<td>NTF + $l_1$ norm ($\mu = 0$, $\lambda = 0.05$)</td>
<td>0.6684 (0.0385)</td>
<td>0.1821 (0.0105)</td>
<td>0.8914 (0.0111)</td>
<td>0.3068 (0.0023)</td>
</tr>
<tr>
<td>NTF + logit ($\mu = 0.1$, $\lambda = 0$)</td>
<td>0.6743 (0.0328)</td>
<td>0.1788 (0.0069)</td>
<td>0.8956 (0.0081)</td>
<td>0.3076 (0.0034)</td>
</tr>
<tr>
<td>NTF + $l_1$ norm + logit ($\mu = 0.1$, $\lambda = 0.05$)</td>
<td>0.6852 (0.0259)</td>
<td>0.1822 (0.0066)</td>
<td>0.8913 (0.0075)</td>
<td>0.3079 (0.0025)</td>
</tr>
</tbody>
</table>

Patterns showing $p$-value < 0.05 were: Pattern 2 for CN and Patterns 4, 18, 28, and 30 for AD. That is, Patterns 4, 18, 28, and 30 were positively related to AD whereas Pattern 2 was negatively related to AD. To verify this relationship, we reported the number of patients from AD and CN according to the extent of involving to each pattern, the membership values (Table 4: Membership distribution). We found that in AD patterns (4, 18, 28, and 30) the ratio of AD patients to CN increases as the membership value increases. In contrast, in CN pattern 2 the ratio of AD patients to CN decreases as the membership value increases. These findings are consistent with the logistic regression results in which AD patients are more likely to have larger values on the Patterns 4, 18, 28, and 30; and to have smaller values on the Pattern 2. Note that we initially made cohort with ratio of AD:CN=1:4, thus the number of CN was always larger than the number of AD across all membership values.

We visualized the five representative patterns (Table 4: Visualization). We represented each pattern using heatmap, according to the membership values of frequency bands or time slots. Frequency bands or time slots that dominantly characterize the pattern showed high values (bright yellow), whereas frequency bands or time slots that are less involved in the pattern showed low values (dark purple). Slow oscillation, slow wave activity, and spindle were overlapped with Delta and Sigma waves, so we denoted them separately where each band belongs to.

Patterns can be interpreted in details as follows:

- **Pattern 2 - Healthy normal sleep.** Pattern 2 was more prevalent to non-AD group. Pattern 2 forms a slow oscillation, slow wave activity and spindle, which none of AD EEG patterns have. Pattern 2 is consistent with the existing known fact that AD patient rarely forms spindle in Sigma wave (12 - 15 Hz) [34].

- **Pattern 4 and 18 – Abnormal REM sleep.** Pattern 4 refers to Delta and Theta wave (0.5 Hz - 8 Hz) with slow oscillation and slow wave activity and also Gamma wave (30 Hz -). This activity occurs during at the beginning and the end of sleep. Similarly, Pattern 18 refers to Theta wave (4 - 8 Hz) with Gamma wave (30 Hz -) during the middle of sleep (4 - 7 hour). Because Gamma wave are usually seen during REM sleep, we hypothesized that this combination of slow wave and high frequency wave is from repeated REM sleep stage.

- **Pattern 28 - drowsy sleep.** Pattern 28 was focused on Theta wave 4-8 Hz and Alpha 8-10.5 Hz waves several times throughout the 8 hours of sleep. This pattern is consistent with typical AD patient’s sleep pattern. It is known that AD patients usually wake up several times during nighttime [37]. As a result, the percentage of wakefulness and Stage 1 increases, which refer to the transition of the brain from Alpha waves (8-10.5 Hz) to Theta waves (4-8 Hz).

- **Pattern 30 - Insomnia.** Pattern 30 has very strong signals on Beta and Gamma waves (15 Hz -) during the first one hour and 5 - 8 hour of sleep. This sleep EEG pattern is accordant with that of insomnia patients, in which Beta and Gamma activity usually increase [38]. Insomnia is one of common comorbidities of AD patients [1].
Table 4. Summary of representative five patterns. Logistic regression coefficient and p-value to classify AD from CN. Membership values distribution of AD and CN, and visualization of each pattern. SO= slow oscillation, SWA=slow wave activity. Range of membership values varies according to pattern. SO, SWA, and spindle were presented separately due to overlap.

<table>
<thead>
<tr>
<th>EEG pattern</th>
<th>Coefficient (p-value)</th>
<th>Membership distribution</th>
<th>Visualization (Dominant frequency bands and time period)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (CN)</td>
<td>-2.7164 (0.034)</td>
<td>AD, CN</td>
<td>SO, SWA, Spindle</td>
</tr>
<tr>
<td>4 (AD)</td>
<td>1.8200 (0.013)</td>
<td>AD, CN</td>
<td>SO, SWA</td>
</tr>
<tr>
<td>18 (AD)</td>
<td>1.4873 (0.034)</td>
<td>AD, CN</td>
<td></td>
</tr>
<tr>
<td>28 (AD)</td>
<td>2.4115 (0.028)</td>
<td>AD, CN</td>
<td></td>
</tr>
<tr>
<td>30 (AD)</td>
<td>2.1408 (0.013)</td>
<td>AD, CN</td>
<td></td>
</tr>
</tbody>
</table>

LIMITATION AND CONCLUSION
Data collected in the National Sleep Research Resource (NSRR) was meant for study the impact of sleep disorders on important health outcomes. Although many clinical variables have been collected, they are mostly focused on sleep but not specific to the AD population. Based on the populational AD occurrence rate (at different ages), we believe there are more eligible patients for the study but it is hard to confirm. So, we end up with a relatively small population but results are still very interesting as we identified several distinctive patterns that are significantly
associated with AD. These results still need further verification and biological explanation. Another limitation is related to the onset time of AD.

In this study, we aimed to derive sleep patterns of AD patients using EEG signals. From the National Sleep Research Resource, we built a unique cohort of 83 cases and 331 controls with clinical variables and EEG signals. We used a regularized nonnegative tensor factorization that can derive discriminative patterns. Among the 30 patterns extracted, we identified five significant patterns with \( p < 0.05 \): health normal sleep with spindle; abnormal REM sleep; drowsy sleep with repeated wakefulness; and insomnia. This study is preliminary but finding are interesting, which is a first step to provide quantifiable evidences to measure sleep as a risk factor of AD.

REFERENCES


HarborBot: A Chatbot for Social Needs Screening

Rafal Kocielnik¹, Elena Agapie¹, Alexander Argyle¹, Dennis T. Hsieh², Kabir Yadav², Breena Taira², Gary Hsieh¹
¹University of Washington, WA; ²UCLA Medical Center, CA

Abstract

Accessing patients’ social needs is a critical challenge at emergency departments (EDs). However, most EDs do not have extra staff to administer screeners, and without personnel administration, response rates are low especially for low health literacy patients. To facilitate engagement with such low health literacy patients, we designed a chatbot - HarborBot for social needs screening. Through a study with 30 participants, where participants took a social needs screener both via a traditional survey platform and HarborBot, we found that the two platforms resulted in comparable data (equivalent in 87% of the responses). We also found that while the high health literate participants preferred the traditional survey platform because of efficiency (allowing participants to proceed at their own pace), the low health literate participants preferred HarborBot as it was more engaging, personal, and more understandable. We conclude with a discussion on the design implications for chatbots for social needs screening.

Introduction

Although designed for medical emergencies, emergency departments (EDs) have become a common place where patients seek help for various problems. EDs thus care for patients who present for not only heart attacks and strokes, but also a variety of social ills, such as homelessness, poverty, and hunger. These social concerns disproportionately affect the poor and limit their access to healthcare. Left unaddressed, social determinants drive up cost and utilization, and are a fundamental source of persistent health disparities.

However, despite the growing body of evidence that addressing social determinants improves health outcomes, assessing patients’ social needs remains a critical challenge. Most EDs currently do not screen for social needs and do not have extra staff to administer screeners without interrupting clinical workflow. Without personnel administration (such as research assistant, nurse, etc.), response rates for both paper and electronic surveys are low. This is compounded by the fact that only 12% of Americans have proficient health literacy.

Chatbots offer multiple potential benefits for social needs screening. Chatbots are systems designed to engage with users through natural language, mimicking a human-to-human interaction. Popular examples of chatbots include Apple’s Siri, Google’s Now, and Microsoft’s Cortana. Extended to the context of social needs assessment chatbot can support self-administering of social needs screeners to minimize personnel cost. In contrast to current form-based survey, a conversational approach would be more “chat” like, potentially offering a sense of familiarity similar to mobile text messaging. By creating a sense of interacting with another person, the chatbots may also increase participation engagement. Furthermore, offering text-to-speech output can also facilitate comprehension.

In this work, we present HarborBot, a chatbot for administering social needs screeners. HarborBot’s persona is designed to be professional, trustworthy, and supportive. It uses a chat-like interface to ask the social needs screener, and allows survey takers to ask for clarifications as if they are interacting with a survey administrator. To study its use, we recruited 30 participants, which includes both high and low health literacy individuals, with both ED and non-ED patients. Through a within-subjects study design, participants were asked to take the survey twice, once via HarborBot, and once via an existing survey platform, SurveyGizmo (order randomized).

Our finding suggests that there is a clear divide in preference between the high health literate (HL) and low health literate (LL) participants. Almost all HL participants preferred taking the survey via SurveyGizmo, whereas the LL participants preferred HarborBot. Post-study interviews reveal that HL individuals consider conversational approach to be too slow without allowing them to take the survey at their own pace. On the other hand, LH participants appreciated the audio output of the HarborBot, and thought the conversational design was much more engaging and personal.

Our work offers numerous insights on the role of conversation-based approach for surveys, and advance our understanding of how to more effectively screen for social needs for both high and low health literacy patients.
Related Work

With the growing interests in clinical screening, research has examined the use of technology-based solutions to support the self-administering of surveys, or Computer-Assisted Self-Interviewing (CASI)\(^2\). CASI include the use of online survey platforms, mobile apps, and electronic kiosks. These self-administered solutions may help maximize scalability and speed of data collection while reducing cost. Compared to face-to-face interviewing, self-administered solutions can also reduce social desirability bias, and limit the under-response in sensitive issues\(^2\). Furthermore, technology-based solutions led to fewer item missing responses, with the same response rates, compared to pen and paper self-administered questionnaires (PAPI)\(^4\).

A variation of CASI is Audio Computer Assisted Self-Interviewing (ACASI). The ACASI system allows users to listen to pre-recorded or text-to-speech audios of questions as if they were being asked by experimenters. This is especially valuable for supporting the understandability among low literacy participants\(^5\). The use of ACASI also reduces social desirability bias, making it an effective format to collect sensitive information\(^6\). This is valuable when assessing social needs in the diverse ED population. One drawback is that ACASI systems take longer to interact with\(^7\).

Despite these advantages of existing technology-based solutions, face-to-face is still better when it comes to response rates (in one study, 92.8% face-to-face response rate compared to 52.2% web-survey response rate)\(^8\). These differences, also in ED context\(^9\), have been linked to the motivating impact of interpersonal interactions, but reproducing such effects via technology is still a challenge. Additionally, non face-to-face surveys often bias against nonwhites\(^10\), low income patients, homeless or those that are disenfranchised with mental health and/or substance use\(^9\).

The Potential Role of Chatbots in Survey Screening. The idea of using an interactive system to collect and deliver health-related information has been studied primarily via so called embodied conversational agents (ECA)\(^11\). ECAs allow people to interact with physical or graphical agents that embody a person in appearance, behavior, and dialect. Mimicking a human-like interaction, ECAs are able to improve engagement and trustworthiness\(^12\). They are also perceived by patients with varying literacy levels as acceptable and easy to use\(^13\). However, their potential for screening is less well studied. Only a few research systems have been prototyped to explore patients interviewing\(^14\), and the actual use of these systems is quite limited, as the cost of development these human-like embodiments is high.

One potential lighter-weight alternative to ECA are chatbots which are composed of six key features\(^15\). First is the concept of thread as app, where the app-centric homescreen is replaced by a threaded conversation with streams of messages and notifications. Much like the interaction one has with another through chat messages. Second is history awareness, where the bot keeps a log of past interactions with users in the thread. Third is an enhanced user interface (UI), where the interaction need not be limited to plain text, but can also include images, audio output, structured messages. Fourth is limited natural language processing (NLP), which is to prevent breakdowns due to technology limitations. Still, chatbots try to mimic conversation with a human partner, e.g., by using “is typing” indicators, dynamic utterances and persona-driven emotional responses. Fifth is message self-consistency, where the intent of each message is clear and stands on its own. Finally, use of guided conversations to prevent users from getting lost.

Like ECAs, chatbots have multiple features that can be ideal for a low literacy population. For example, the audio output can facilitate understandability; the conversation-like interactions can foster a sense of interacting with another person, making the interaction more personal and engaging. But chatbots also differ from ECA in a couple of critical ways. One is that chatbots do not require a graphical or physical embodiment of the agent. This minimizes the cost of development and offers a more scalable solution. People can interact with chatbots even through their mobile devices. Another key difference is that chatbot mimic text messaging. Demographic surveys have shown that people with lower SES (more likely to have lower health literacy and higher social needs) are increasingly reliant on mobile text messages to communicate\(^16\). The chat-like interface may be more welcoming and intuitive for those users.

Chatbot based surveys were shown to increase user engagement and produce higher quality responses\(^17\). Such effects, however, have only been demonstrated with low-stakes demographics and marketing online surveys, and with general population. It is unclear if and how such benefits could be translated to clinical setting and low literacy population. Furthermore, low literacy in ER poses unique challenges in term of understandability and comfort with sharing information. Finally, prior work has provided very little in terms of linking the particular chatbot design aspects to their effect on the users, which limits its value for informing future design decisions. Our work aims to fill these gaps.
We designed and implemented a custom chatbot called HarborBot to test a conversational approach to surveys. HarborBot interacted with users through chat and voice. It communicates via chat messages, that it can also read out, as if it is speaking. Users interacted with the system primarily through buttons (for structured responses) and text (for text-based questions). HarborBot is implemented as a webapp and we had our participants interact with it on tablets.

**Design Process.** To create HarborBot we followed an iterative design process in which a team of 2 senior HCI researchers and 6 design students followed three general design phases: 1) Requirements gathering - the team consulted 3 ED practitioners, who are also co-authors on the paper, and existing literature related to patient experience in ER, 2) Design exploration - the team explored various low-fidelity prototypes and gathered feedback on them from ED practitioners and via small scale usability tests, 3) Refinement - the most promising prototype was developed further and refined with positive elements from other prototypes. This process resulted in the final HarborBot system.

**User Interface.** We used BotUI - a Javascript framework to build conversational UIs. Messages to and from HarborBot appear in standard elliptic chat bubbles with users messages distinguished from the bot’s by different colors. Prior to the appearance of messages from HarborBot, animated ellipses are shown in the chat bubble with a delay to denote that the bot is typing (Fig 1 f), akin to that of iMessage or similar interfaces.

BotUI allows for the creation of different question types, which we used to cover the types of questions asked by the screener. Each message from BotUI would be one of these types: skip, yes/no, input, options, or many options. Skip (see Fig 1 c) was a unique type that would move onto the next message of HarbotBot’s script without user responding. Yes/no (Fig 1 A), and options (Fig 1 B) use the standard buttons offered by BotUI. For questions involving multiple possible answers (i.e., checkboxes), we used a vertically stacked list of options that allowed users to choose multiple options before submitting (Fig 1 D). Free response (Fig 1 C) required a text field and device’s keyboard to respond.

**Persona.** There were several personality considerations we made in the design of HarborBot. Most importantly we emphasized striking a balance between a serious and friendly tone. If HarborBot was too friendly, users may feel the conversation is not being taken seriously. If too serious, users will lose the feeling of comfort we want them to have in answering personal questions. We also intentionally avoided any use of humor seeing as it would be inappropriate in the context. In addition, given the stressful nature of patients’ experiences in EDs, we sought to make HarborBot empathetic. However, we tried to do so without pitying the user or being condescending in any way. To accomplish this, HarborBot used occasional confirmatory phrases, such as: “Okay, I’m getting a better idea of where you are at.”, “Got it”, and assurances, such as: “The next questions are about your personal safety and may be tough to answer.”

The voice of HarborBot was an important part of the interaction with the user. By default, HarborBot used a female

---

1BotUI - https://botui.org/
voice taken from the Microsoft’s Bing Voices\(^2\). Users could adjust the volume of the voice or mute it entirely for privacy reasons or personal preference.

**Dialogue-Based Interactions.** We employed thread as app principle to support survey-taking. Survey questions and user replies were presented as streams of messages in threaded conversation akin to chat messaging. We allowed users to skip and ask for clarification of the questions. Each question posed by HarborBot offered a couple of utility options using round buttons next to the answer area. The Skip button (Fig 1 c) allowed the users to skip questions that felt uncomfortable or they couldn’t answer. After skipping, HarborBot would proceed on with the script. We supported rephrasing the question to offer its simplified version for low literacy individuals (Fig 1 b). We used Readable to make sure our rephrased versions were at most at a fifth grade reading level. Selecting this option is akin to saying “I didn’t quite understand that.”, which would make HarborBot re-ask the question using the simplified phrasing.

Additionally, we implemented an edit button (Fig 1 a) next to each past answer in case user needed to change it. Selecting this option would make HarborBot ask the corresponding question again and allow the user to provide a new answer. This is akin to users saying to the bot “Can I change my response to that question?”

Occasionally, HarborBot would respond with conversational remarks. These utterances were essential to developing Harbor’s personality, and engaging users in a conversation. Some of these interactions are dynamic based on a rule-based approach. For instance, if a user indicated they did not have a steady place to live, HarborBot would not ask the remaining housing questions. If the user response indicated a negative social situation, HarborBot would acknowledge it with a sympathetic affirmation, such as “That must be stressful, I’m sorry to hear that.”

**Method: Study**

We conducted a within-subjects study with 30 participants to compare the experience of answering a social needs survey using two different platforms: HarborBot (Chatbot) and a more traditional interface for taking surveys - Surveygizmo (Survey). We recruited participants with high and low health literacy at two study sites. We expected the Chatbot interface will be 1) more engaging, 2) more understandable, and 3) more comfortable to share information with, while 4) preserving response quality. We also expected these effects to be pronounced with low literacy users.

**Study Procedure.** Users interacted with both survey interfaces using a tablet’s web browser. After interacting with one interface, participants reported their perceptions and experience. They then repeated the same procedure for the second interface. We randomized the order of interaction. After completing both, we conducted an interview.

**Social Needs Survey.** In both platforms users answered the social needs survey developed by the Los Angeles County Health Agency (LACHA)\(^18\). This survey comes from over two years of work from the committee on the Social and Behavioral Determinants of Health, with members from the Departments of Health Services, Mental Health, and Public Health. The survey asks 36 questions related to demographics, financial situation, employment, education, housing, food, and utilities as well as questions related to physical safety, access to care, and legal needs. A number of questions can be considered sensitive, such as: “Have you ever been pressured or forced to have sex?”, “Are you scared of being hurt by your house?”, “Did you skip medications in the last year to save money?”

**Measures.** Participants evaluated both survey platforms in terms of workload (NASA TLX survey\(^19\)), engagement in the task (questions adapted from O’Brian’s engagement survey\(^20\), e.g., “I was really drawn into answering questions.”, “This experience of answering questions was fun.”, “I was absorbed in answering questions.”), understandability of content (“I understood the questions that were asked of me.”), and willingness to share information (“I was comfortable answering the questions.”). These measures have been commonly used in prior studies of chatbots\(^13\).

Participants health literacy was measured using Rapid Estimate of Adult Health Literacy (REALM)\(^21\) which assesses participant’s ability to read health materials and instructions, at a comprehension level of high school or lower\(^21\). We used the Newest Vital Sign (NVS) health literacy scale\(^22\) to assess likelihood of limited health literacy based on numeracy, prose and document literacy measures. NVS was used for fast recruitment (under 3 minutes).

During interview we asked about preferences for the two survey platform, the specific features of the platforms, participants’ comfort in sharing information in each platform, and perceptions of the personality of the chatbot.

\(^2\)https://docs.microsoft.com/en-us/azure/cognitive-services/speech/api-reference-rest/bingvoiceoutput
Recruitment. We recruited participants from two study sites. In the Seattle metropolitan area participants were recruited through Craigslist, flyers at local community centers, and the Institute of Translational Health Sciences’s research subjects’ pool. In the Los Angeles County participants were recruited from a large county safety net hospital (Harbor-UCLA), by two of the authors who are physicians there. ED visitors were handed out a flier at discharge, or while waiting. Participants were 18 or older and had a conversational level of English proficiency. The study was approved by the IRB at both sites.

Participants. 30 participants were recruited (17 males, 10 females, 3 declined to answer) ranging from 23 to 65 years of age (M=39.63, SD=12.91). They reported completing 13.15 (SD=3.73) years of education on average. 22 reported English and 4 Spanish as their primary language. One person reported bilingual fluency and 3 people declined to answer. Finally we had a diverse ethnic backgrounds: Hispanic or Latino: 9, Black or African American: 8, White: 6, Multi race: 2, and 4 reported other ethnicity or declined to answer.

11 participants were assessed as low, and 19 as high health literacy. Participants were considered low literacy if they scored at a seventh to eighth grade level or below, on the REALM scale, or got a score that suggests high likelihood (50% or more) of limited literacy on the NVS scale.

Analysis. Quantitative data comprised user responses for Chatbot and Online Survey interface, Chatbot interaction logs, and post-interaction survey responses for each platform. These were matched by participants’ unique id and the analysis focused on descriptive statistics of user interactions, especially with Chatbot, and on comparison of answer equivalence for the two platforms. Differences in survey responses were assessed using paired t-tests and interactions between interface type and participant’s health literacy levels were explored using linear mixed effects models.

The interviews took between 7 and 25 minutes (M=17.56, SD=9.21), conducted by three and analyzed by four of the authors. Each researcher wrote a detailed summary of interviews they had not conducted, including quotes. We then developed a codebook following a top-down and bottom-up approaches. Initial codes for the top-down pass were informed by the interview questions (why participants liked or disliked each survey platform, attitudes towards sharing information, perceptions of Chatbot). The interview structure itself was informed by the literature. We then refined the codes based on themes that emerged from the data in a bottom-up fashion. Each interview summary was coded by a researcher on the team (who had not conducted the interview, or written the summary). The coded interview summaries were used to identify themes. Three of the authors discussed the overall themes until consensus was reached. Researchers consulted with the audio and transcriptions of the interviews to ensure validity of the coding.

Quantitative Results

Preferences. Low health literacy (LL) participants preferred using Chatbot over the Survey with 8 out of 11 expressing such preference. At the same time, 17 out of 19 high literacy (HL) participants preferred Survey. This difference was statistically significant ($\chi^2 (2, N = 30) = 12.5$, p < .001).

Time to Completion. Participants had to respond to 36 questions in the social needs survey, but they could also skip answers. They spent significantly ($t(27)=2.23$, p<0.05) more time answering questions via Chatbot (M=9:26 min; SD=3:14 min) than via Survey (M=6:48 min; SD=6:28 min). We found no significant difference between answering time (avg. of both interfaces) for LL (M=9.43 min, SD=3.23) and HL participants (M=7.36 min, SD=4.20). We also found no significant interaction between the interface and literacy level on time.

Equivalence of Responses. An important question is whether the two interfaces result in the same data quality. We explore two measures: per-item response rates and data equivalence. On average participants provided almost identical number of answers via the two interfaces: 32.93 (SD=3.48) questions answered with Chatbot and 33.00 (SD=2.95) with Survey. This suggests comparable response rates. In terms of data equivalence 87.0% (SD=11.6%) of the responses per user were the same across the two interface versions.

Reasons for Response Discrepancies. We found that skipping an answer in one interface, but not the other was the primary cause of answer discrepancy (48% of mismatches). There was, however, no significant difference between the two platforms in skipping behaviors. 25% of mismatches was a result of skipping a question in Chatbot only and another 23% due to the opposite. Furthermore, the order in which users encountered the interfaces had no significant impact on skip rates: 8.0% (SD=9.3%) when answering the survey the first time, and 7.8% (SD=8.2%) when answering
the survey the second time. Hence the platforms are not different in this respect. One interesting finding from our explorations is that there seems to be an anchoring effect with users skipping more often when starting the study with Chatbot, for their responses to both platforms: Chatbot (M=9.8%, SD=29.7%) and Survey (M=9.8%, SD=27.2%) than when starting with Survey: Chatbot (M=5.2%, SD=22.3%) and Survey (M=4.9%, SD=21.6%). This is most likely due to the skip option being more explicit in the Chatbot and users wanting to be consistent in their answers.

Manual examination of the remaining mismatches revealed varied and non-systematic reasons for discrepancies such as: low equivalence only in the very first introductory question (53.3%), direct contradiction (e.g., user answered “Yes” in one interface and “No” in the other); similar, but not the exact same answers (e.g., answer: “Yes, help finding work” vs. “Yes, help keeping work”), ticking an additional option in a multi-choice answer (e.g., “Unemployed - looking for work” vs. “Unemployed - looking for work, Disabled”) and a possible misinterpretation of the question (e.g., when asked for income per month, user typed “2000” in one interface and “24,000” in the other).

Workload (NASA TLX). Analysis of the NASA TLX survey responses revealed a difference in task load index (avg. of all items denoting workload, α: 0.83) between Chatbot and Survey. Participants reported a higher workload when using Chatbot (M=2.460, SD=1.241), compared to Survey (M=2.167, SD=1.284; t(27)=-2.020, p=0.05). Given the scale from 1–lowest to 7–highest, this still represents a low perceived workload. We also found a main effect of literacy level: there was a higher perception of workload across both platforms by the LL participants (M=2.955, SD=1.335) than the HL ones (M=1.921, SD=0.948; t(27)=2.439, p<0.05). The interaction effect was not significant.

Engagement, Understandability, and Comfort with Sharing Information. Analysis of the engagement index (average of O’Brian’s engagement questions, α: 0.82), revealed a higher reported engagement for LL participants (M=3.920, SD=0.502) than HL ones (M=3.469, SD=0.402), (t(27)=2.672, p<0.05). We also found a weakly significant interaction between interface and literacy with LL participants being more engaged with the Chatbot than HL ones, but less engaged with the Survey (Chatbot*Low, β=0.485, SE=0.262, p=0.064). This represents a half a point increase on a 5-point likert scale for engagement. Trends in the same direction, but no significant differences were found for understandably and comfort with sharing information.

Qualitative Results
In this section, following mixed-methods approach, we complement and expand on the quantitative findings. Participants varied not only in their preferences for Chatbot or Survey, but also in the particular aspects they liked about each, as well as in which design aspects were instrumental in creating particular perceptions and experiences. Participants valued the engaging conversational aspects of the Chatbot. Especially LL participants found the conversational interface more caring in the context of a sensitive topic. In contrast, HL valued the efficiency of the SurveyGizmo interface and felt slowed down by the Chatbot. Some participants found the Chatbot more robotic, disingenuous or pushy at times, but these seem to result from the particular way in which HarborBot implemented conversation.

Strengths of Our Conversational Approach
Engaging. Most participants found the conversational features of the chat more engaging than the Survey, regardless of the health literacy level. Participants felt like they were having a conversation with a person when using the Chatbot. More than half the participants attributed such perception to the use of voice: “she was reading the questions and I can answer it ... seemed like a conversation ... like someone was talking to me and it gave me the opportunity to answer back and then they answered back” (H59). Other participants felt the ellipses made it feel like having a chat with someone (H76, L77), and even referred to the the Chatbot as “she” (8 participants). Some participants valued that the Chatbot felt like a person: “I liked... how it talked to you, reads you the questions ... it spoke directly at me” (L60), “I thought it was someone asking me those questions” (L72).

Aside from the voice and ellipses, the conversational utterances also contributed to the perception of interacting with a person (L75, L58, L72, H32, L60, L36, H41, H59). One participant found them motivating: “Saying 'you got it,' It’s giving you motivation ... nice to hear that once in a while” (H73). Another felt like the conversation was adapting to the answers to be more relevant: “seem like they tried to give you a little positiveness based on your answer” (H59).

Caring. Participants perceived the Chatbot as caring, particularly in the LL group. These participants had a generally positive attitude towards the social needs survey questions (L51, L55, L58, H73) and this topic resonated with their
personal experiences “It felt like it was telling me about my life. That was really amazing, like woow” (L71). Therefore, some of the perceptions of the Chatbot might have been accentuated by the positive perception towards the survey topic. Many participants described the personality of the Chatbot using terms such as: caring, kind, patient, helpful, calm, familiar, or concerned (H35, H41, L52, L55, L57, L61, L77). Participant also reported the voice of the Chatbot was aligned with this caring personality: it was soothing (H57), had cadence (H32), helped a nervous participant feel more comfortable (L55) and was “nice and sweet made me feel relaxed” (L77).

The Chatbot was designed to provide supportive utterances in response to some of the participants answers. Many participants liked these utterances (L60, L36, H32, H41, H58, H59). One participant though the utterances made him feel “comfortable to answer the questions” (L61), and that they provided a positive reinforcement to keep on answering (H59). Participants perceived Chatbot utterances such as “I am sorry to hear this” as the Chatbot “trying to be understanding” (H59). Some found these utterances to be very applicable to the conversation context. For example L61 considered the Chatbot response: “That must be stressful” to be a reaction to the information she shared: “she probably said that because of my financial situation” (L61), which she felt would be calming for people “to not be stressed, I would think it would be helpful” (L61). Other participants felt the supportive utterances gave them confidence: “nice lady giving me confidence ... with good tone of voice” (L75).

Understandable. Several LL participants (5 of 11) reported having trouble with reading and understanding the written questions in the Survey. They liked using the Chatbot because it facilitated their understanding, which they attributed to the audio feature: “When I hear it I have a better understanding of the question” (L61) or that “just hearing it I could ... relate better to the question” (L53). Some participants reported using the feature that replayed audio, to better understand a particular item (L51, L58, L61, L73). This was especially useful when they missed some words or did not fully comprehend some of the contents at first: “I didn’t get it at first, so I wanted to go back and listen to it again before answering” (L58). Several also mentioned that they would have liked it if the answers were spoken via audio as well, to make them more understandable (L61, L54).

Accessible. Some participants had particular needs that the Chatbot was able to satisfy much better than the Survey. One participant who reported vision problems, preferred having questions read to them: “If it is too small I can’t see it so I prefer to have the questions read to me anyways” (H73). Another participant reported feeling very comfortable with the Chatbot because she was regularly experiencing panic attacks and considered ED stressful: “I was thinking I was texting somebody ... that made me forget where I was at ... it was like texting my sister my mom and waiting for them to respond back. And that made me feel patient” (L77). In contrast, she found it particularly difficult to take the Survey: “by myself ... it felt awkward and alone” (L77).

Weaknesses of Our Conversational Approach

Inefficient. HL participants cared about efficiency, primarily reflected in the speed of completing the survey. The majority of HL participants (17 out of 19) preferred to use the Survey because of that. Several mentioned that the traditional interface enabled them to be faster than the Chatbot (H21, H22, H24, H59), or to go at their own pace (H36). Participants attributed being slowed down to various conversational features of the Chatbot. Some felt the Chatbot was slower because they needed to wait for the ellipses before a new question would appear (H35). They were also able to read faster than the questions were read by the bot: “when she was talking at me. I felt like I was going at a slower pace” (H23). Also not having to engage with additional conversational utterances was seen as more efficient (H35, H56). The audio feature was perceived as interfering with reading and thinking (H23, H40, H70, H21, L71). One LL participant preferred the Survey because they could concentrate more: “to read is better ... Because that way I could like concentrate more and think about more and you know ... I could read my letters more and makes it better for me.” (L58).

Pushy. Somewhat surprisingly, a few participants perceived Chatbot as being pushy, based on the tone and the speed at which questions were asked. Some participants felt the questions asked were very direct (H57, L72, H52). L72 felt like he was answering questions to a teacher, and had to provide correct answers. H57 and L72 thought there could be more utterances to help prepare the survey taker for some very sensitive questions in the survey. H57 also felt that some of the questions were trying to repeatedly get information that he had already declined to provide: “if I say none of the above ... don’t be pushy” (H57). Others also felt rushed in providing the answers to the Chatbot. For example,
the use of ellipses, and the short delay between its messages made it feel like the Chatbot was moving faster than the participants were comfortable with (H23, H63). Participant H63 felt like the questions kept coming and he had no control over when they would be read.

**Robotic and Disingenuous Voice.** Some participants, primarily in the HL group, perceived the Chatbot as being robotic. Some participants found the voice not sounding natural (H21, H22, H23, L58, H59, H63, H70, H76), for example sounding “truncated...monotone...seemed pretty artificial to me.” (H70). Some perceived the Chatbot as disingenuous when the utterances did not meet their intended purpose (H63, H40, H23, H52): “I feel like they were trying [to make] the software to feel sympathetic, or empathetic, that was weird” (H63). Another participant perceived utterances as defaults: “it felt like defaults rather than someone 'feeling for you’” (H40). The perception of artifical responses led another participant to perceive the Chatbot as fake, and was reminded of customer support: “kind of just programmed, recorded in, to appear to be more personal...hell there’s nobody there somewhat disingenuous... It reminded of... dealing with the phone company” (H70).

**Inconclusive Impact on Willingness to Disclose Information.** Most participants, regardless of health literacy level, reported being comfortable sharing information asked by the survey questions. However, the human-like interactions of Chatbot did affect some participants’ willingness to disclose information, although participants reported effects in both directions. For some, if they thought they were interacting with a person, they felt more reluctant to share sensitive information, or tell the truth: “I might be more honest if I’m reading [the question] ... if someone else ask me about them, I might lie” (L72). Another participant showed concern about the identity of the potential conversational partner: “it was a robot, I didn’t mind, but I think if it was a human being I would mind... and you really don’t know who’s on the other end” (H40). In contrast, some participants were more willing to disclose because of the human-like interactions. “If it says ‘I would like to more about you’. It gives me the confidence to open up, because each question that follow sounds so interesting and it gives me the opportunity to interact with the person on the other side ... it wetens my appetite to give out more information” (L75).

**Discussion**

**Main Findings.** In this paper, we proposed the use of a chatbot (HarborBot) for social needs screening at emergency departments and compared it to a traditional survey tool (SurveyGizmo). Based on interviews, interaction logs, and survey responses we demonstrate that the conversational approach is perceived as more engaging by all the participants, and further as more caring, understandable, and accessible among the low health literacy (LL) ones. Importantly, we also demonstrate that the conversational approach results in similar response rates and 87% equivalence in the collected data. At the same time, we found the conversational approach to be more time consuming (in line with reports from prior work on ACASI) and prone to be perceived as somewhat pushy, robotic, and disingenuous which was, however, mostly the perception of participants with high health literacy (HL).

**Positive Design Aspects.** Numerous strengths of the conversational approach for LL population can be linked to conversational features. First, various features of the chatbot facilitate understanding. The audio output is especially valuable for participants who are less proficient readers. Second, the ability to ask the bot to rephrase the question offered a way to ask for clarification that is currently not a feature in online survey platforms. Third, chatbots can create a sense of interacting with a human. The utterances can make the survey takers feel cared for and engaged. Such positive interactions made some participants feel relaxed and even motivated to answer more questions.

**Challenging Design Aspects.** We were surprised that the conversational features felt pushy for some, especially HL participants. Such perception was linked to the tone of the questions and to the speed of the interaction. In terms of tone it is possible that our literal use of the wording of the survey questions was not the most appropriate for creating a conversational feel. In terms of speed of interaction, the use of voice might be a contributing factor. As reported in prior work agent asking questions via voice can create a perception of response urgency. This could be improved by adding assurances like “please take your time.”, manipulating intonation, or making it more explicit that the ellipses represent someone is typing (rather than the system is waiting for a response). The second reason for pushy feel could be related to the fixed speed of conversation. Human-human conversation involves not only exchanging information, but also coordinating various aspects of the exchange, e.g., its speed. If a participant needs more time to think, a real person, would pick it up from verbal and non-verbal cues and adjust the speed. Our HarborBot is currently incapable
of making such adjustments. Such fixed speed may feel too fast or too slow for some users.

HarbotBot felt “caring” for LL and “robotic” for HL participants. This might be related to the different expectations and tolerance levels for voice quality and may be improved with use of a better quality text-to-speech service (technical challenge), human pre-recorded audio clips (which comes with limitations in flexibility), or modifications of intonation and prosody using approaches such as Speech Synthesis Markup\(^3\). Another way may be to generate more personalized and diverse utterances\(^24\).

**Future Design Directions.** Given the division of preferences for chatbot/survey between the HL and LL groups, one possibility for a real-world use could be to have two versions of the tool and either intelligently assign or have patients pick the version they would prefer to engage with. While, long waits in healthcare setting make it less of a problem, a number of design opportunities can still be explored to make the chatbot interactions more efficient, such as simplifying the script, or providing user control over time between messages. While we focused on examining the effects of the conversational approach for a LL population, our findings suggest a potential for accessibility-focused uses of the chatbot. Participants who were hard of seeing mentioned they appreciated the audio output. Further, one participant with anxiety attacks appreciated the human-like interactions, which made them feel like chatting with a loved one, at home.

Finally, it is not clear based on our results, how the conversational approach affects people’s comfort in responding to questions, and any potential desirability biases. Prior work suggests that the self-administered screeners would reduce social desirability bias, and limit the under-response in sensitive issues\(^6\). This is because people will not feel like someone is monitoring or judging them. We thought the Chatbot may strike a happy medium between being perceived as human-like to enhance engagement, while not being perceived as a person for people to feel uncomfortable with disclosures. It is not clear if we were able to achieve that balance. Some participants who thought the Chatbot was human-like did not mind sharing and commented that it was more motivating, while others that thought the Chatbot was human-like were concerned with sharing. It is possible that the very initial greeting from the bot sets the tone for the rest of the interaction\(^25\). This requires additional research.

**Limitations**

Recruiting low health literacy participants with basic English proficiency is difficult. Thus our sample size for this group is limited. Still, our results appear fairly robust; our claims are supported through both quantitative and qualitative data. With our quantitative data, even with the small sample size, our primary outcome variables achieved statistical significance. Also with our qualitative results, we reached data saturation early.

We were able to study actual ED patients interacting with HarborBot, which boosts the ecological validity of our findings. Nonetheless, a number of important generalizability questions require future work. For example, would our general findings hold with a different population (e.g., non-English speakers, different cultural backgrounds), or a different set of questions? Furthermore, while we focus on examining the acceptability of chatbot for social needs screening in the ED, a critical next step is to study its feasibility and how it may be integrated into existing workflows.

**Conclusion**

In this paper we have proposed the use of a chatbot as a tool for social needs screening in emergency departments. We designed and built HarborBot to enable a more human-like interactions during the self-administering of surveys. Through a mixed-methods study with 30 low and high health literacy participants recruited at two different sites, we showed that compared to traditional online survey, the conversational interface offers benefits such as increased engagement, facilitating understandability and making the interaction feel more personal and caring. These benefits were especially appreciated by low health literacy participants. However, the high health literacy ones mostly preferred traditional survey as it was more efficient. Our work advances the understanding of conversational agents, and offers valuable insights on the design and potential role of chat interfaces for social needs screening.

\(^3\)https://www.w3.org/TR/speech-synthesis11/
References

Balancing Functionality versus Portability for SMART on FHIR Applications: Case Study for a Neonatal Bilirubin Management Application

Polina Kukhareva, PhD, MPH\textsuperscript{1}, Phillip Warner, MS,\textsuperscript{1} Salvador Rodriguez, PhD\textsuperscript{1}, Heidi Kramer, PhD\textsuperscript{1}, Charlene Weir, PhD, RN, Claude Nanjo, MPH,\textsuperscript{1} David Shields, BA,\textsuperscript{1} Kensaku Kawamoto, MD, PhD, MHS\textsuperscript{1}

\textsuperscript{1}Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

Abstract

**SMART on FHIR applications** are standards-based tools integrated with electronic health record (EHR) systems and intended for dissemination across EHR platforms. A key challenge for disseminating many apps is that EHR vendors provide different levels of support for FHIR. Thus, app developers must balance functionality versus portability. In this case study, a feature-rich app for neonatal bilirubin management was developed prioritizing physician-requested functionality, with custom FHIR interfaces implemented within the EHR as needed. Following wide intra-institutional use, several approaches are being pursued for adapting the app for cross-institutional dissemination: user surveys and interviews to identify least-valued app features which could potentially be omitted; enabling the application to provide differential features depending on available EHR FHIR capabilities; replacing custom FHIR interfaces with native EHR FHIR interfaces as they became available; and using a canonical logical data model known as *QUICK* that can be mapped to different FHIR versions and profiles.

**Introduction**

**SMART on FHIR applications**

The Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR)\textsuperscript{1} data interface standard is used by the Substitutable Medical Applications and Reusable Technologies (SMART) platform launched in 2010 for clinical app integration into electronic health record (EHR) systems.\textsuperscript{2} SMART on FHIR applications retrieve relevant data through application programming interfaces (APIs) and are integrated with EHR authorization mechanisms so that they can be embedded within EHR systems with single sign-on capabilities.\textsuperscript{3} Apps can be authored by third parties or by internal app development groups. One of the main promises of SMART on FHIR is its potential for “plug-and-play” interoperability, which could allow an organization to import an app developed elsewhere and utilize it with ease.\textsuperscript{4} Consequently, it has been proposed that SMART on FHIR applications could enable advanced functionality to be cost-effectively disseminated, even to health systems with limited information technology resources.\textsuperscript{5} Every year, more and more SMART on FHIR applications are being developed: the SMART on FHIR app gallery contains 69 SMART on FHIR applications as of March 2019.\textsuperscript{6} Duke Health shared their SMART on FHIR platform information and demonstrated feasibility of deploying 5 apps.\textsuperscript{7} However, the healthcare community is still early in its adoption, implementation, and evaluation of SMART on FHIR applications. Clinical outcomes have not been evaluated yet for most apps and many published reports have focused only on implementation feasibility.\textsuperscript{8,9} Only a few apps have been reported to be used by multiple healthcare systems successfully.\textsuperscript{10}

**Key challenge to dissemination: limitations in EHR-supported FHIR capabilities**

Dissemination is the process of spreading an application from the organization where it was developed to multiple EHRs. We define portability in this context as how easily an application can be disseminated. EHRs are generally compliant with the US Core FHIR profiles,\textsuperscript{11} which define a minimum set of data interfaces using the FHIR standard. However, these FHIR profiles have limitations. For example, the official version of these profiles as of March 2019 (version 1.0.1) does not include support for the FHIR Encounter and ServiceRequest resources, retrieval of clinical notes, or filtering a search for past medications based on a date. Even in the most current draft version of these profiles (version 2.1.0)\textsuperscript{12}, there is no support provided for writing data into the EHR, placing orders, or retrieving relevant clinical data such as family health history, baby’s birth time, or detailed information on a patient’s smoking history (e.g., packs per day). Moreover, even when included in the US Core FHIR profiles, terminologies may not be standardized in the profiles (e.g., medication routes are not required to use a standard terminology), and EHRs may differ in their interpretation and implementations of the profiles. Differences may also exist at the level of healthcare systems using the same EHR platform, providing further challenges to dissemination. Thus, to maximize the potential for dissemination, SMART on FHIR applications must constrain their data...
requirements to those supported by US Core FHIR profiles, and in particular to those profiles that are most uniformly implemented by EHR vendors. While such an approach may not be an issue in some cases (e.g., when developing an application for demonstration or proof-of-concept purposes, or if a use case can be fully satisfied using available data interfaces), in many cases the desired user functionality may not be supportable using existing EHR FHIR interfaces. This situation is expected to improve in future years as the US Core FHIR profiles continue to evolve, and as EHR vendors continue to increase their FHIR capabilities. However, for the foreseeable future, the need to balance portability with functionality will be a real concern for SMART on FHIR app developers.

At the University of Utah, a number of SMART on FHIR applications have been developed and implemented for operational clinical use since 2017. The first of these applications to be deployed for production clinical use was a neonatal bilirubin application. This application is in near-universal clinical use in the newborn nursery, and it is being prepared for external dissemination. Here, we use this neonatal bilirubin application as a case study for examining potential approaches to balancing functionality versus portability of SMART on FHIR applications which can often pose competing demands. The objective of this paper is to assist others developing SMART on FHIR applications to make more informed decisions on how to best achieve the right balance between functionality and portability.

Methods

Implementation Setting

Development of SMART on FHIR applications was performed by the Knowledge Management and Mobilization (KMM) group at University of Utah Health, an academic health system, as part of the ReImagine EHR initiative. University of Utah Health uses the Epic EHR. The KMM team consists of 9 clinical informaticists with expertise in areas including software development, software architecture, standards-based interoperability, and biostatistics. Two of the team members are physicians, 4 are PhDs, 2 are co-chairs of HL7 Work Groups, and 7 are certified to make custom Epic Web Services in the Epic EHR. Team members meet regularly with EHR vendor colleagues to help ensure that any custom interfaces developed are aligned with the vendor’s approach to FHIR. Team members gained expertise in making FHIR extensions to the EHR through the development and implementation of a number of SMART on FHIR apps for clinical use at University of Utah Health, engagement in the standards development process, and technical validations of cross-institutional app deployments in different EHR platforms through inter-institutional collaborations as well as HL7 FHIR Connect-a-thons. The Department of Biomedical Informatics at the University of Utah also has a sociotechnical team which provides evaluation services.

Preparing for Cross-institutional Dissemination

We considered the following approaches for adapting the app for cross-institutional dissemination: (1) initial implementation based on clinician needs, with custom FHIR services implemented as needed, (2) gathering feedback from app users to identify least-valued app features which could potentially be omitted if difficult to disseminate, (3) enabling differential features based on EHR FHIR capabilities, (4) progressive replacement of custom FHIR services with native EHR FHIR APIs, and (5) use of a canonical logical data model known as QUICK that can be mapped to different FHIR versions and profiles.

Initial Implementation Based on Clinician Needs

In 2016, a SMART on FHIR app for neonatal bilirubin management, which we call the Bili App, was requested by pediatric physicians. The KMM team evaluated whether it would be possible to simply use the SMART on FHIR bilirubin app available in the SMART on FHIR gallery as a demonstration application. However, the physicians determined that the original app would not meet their clinical needs and would not be useful to them, due to inadequate support for key recommendations in the underlying clinical guideline from the American Academy of Pediatrics. While the demonstration application required only limited data, such as laboratory data conformant with US Core FHIR profiles, it did not provide sufficient perceived utility for the end users. Thus, the KMM team decided to enhance the app to support the requested functionality. These enhancements impacted portability of the application by requiring data APIs that are not currently supported by US Core FHIR profiles.

Following significant enhancements by University of Utah Health, the Bili App was released for clinical use on April 12, 2017. The app uses FHIR version STU3 (3.0.1), US Core FHIR profiles version 1.0.1 and QI-Core FHIR profiles version 2.1.0. The app was recognized with several awards in the Department of Health and Human
Services’ Provider User Experience Challenge. The app is in near-universal use for babies born in the University of Utah newborn nursery.

Gathering Feedback from App Users

As some of the features were known to be potentially difficult to disseminate on other EHR platforms due to the lack of needed FHIR interfaces in these platforms, we gathered user feedback to determine the most valuable features. Gathering feedback from the users included interviews and surveys conducted by the KMM team’s Director of Evaluation (PK) and a member of the Department of Biomedical Informatics’ Sociotechnical Service Line (HK). We asked medical directors of the newborn nursery and outpatient clinics to invite attending and resident physicians to participate in in-person sessions that included a survey and a semi-structured interview. Three residents and 4 attending physicians agreed to participate, representing both inpatient and outpatient settings. Usability assessment was approved by the Institutional Review Board (IRB).

During the in-person sessions, we showed participants a de-identified screenshot of the Bili App (Figure 1) with 16 screen components: 1) patient bilirubin test results, 2) phototherapy thresholds (green lines), 3) exchange transfusion thresholds (red lines), 4) phototherapy administration (yellow bar), 5) discharge from the hospital (light blue bar), 6) outpatient phototherapy order (orange bar), 7) gestational age, 8) direct Coombs results, 9) mother’s lab results, 10) other neurotoxicity risk factors, 11) albumin < 3.0 g/dL, 12) recommendations (blue box), 13) rebound hyperbilirubinemia risk (green box), 14) table of bilirubin measurements, 15) table of albumin measurements, and 16) hyperbilirubinemia risk tab. First, we asked users to assess the value of all 16 features. Second, we asked whether the providers would still use the app if it no longer included any one of 8 features deemed to be most difficult to disseminate. Responses were arranged according to the median value and visualized using box-plots. Interview times averaged 20 minutes and included screen recordings of clinicians using the app for the assessment of bilirubin levels for providers’ recent actual patients. The semi-structured interviews included the following questions: What decisions are supported by the app? What works well in the Bili App? What are issues and/or limitations? Can you think of any enhancements that would help you?

Enabling Differential Features Based on EHR FHIR capabilities

In the course of other projects that involved integration of FHIR applications across institutions and EHR vendor platforms, the project team came to realize that other EHR vendors may simply not enable third parties to add extensions to their FHIR interfaces. Moreover, through the review of EHR vendors’ FHIR documentation, as well as actual interaction with them through HL7 FHIR Connect-a-thons, it became quickly apparent that EHR vendors differ in their support for FHIR versions and profiles. Thus, we explored potential options for supporting different levels of data availability.

Progressive Replacement of Custom FHIR Services with Native EHR FHIR APIs

During the course of the project, the EHR vendor incrementally added additional FHIR interfaces. Several of these interfaces were capable of replacing the data pulled by custom FHIR services. Also, in discussions with the EHR vendor, it was learned that some of the data points which had been pulled through custom FHIR services could be pulled instead through vendor-provided FHIR APIs following data mapping that could be configured in the EHR.

Use of a Canonical Logical Data Model

In an effort to extend the expressivity of the US Core FHIR profiles to support clinical quality improvement efforts, the HL7 Clinical Decision Support (CDS) and Clinical Quality Information (CQI) working groups developed QI-Core FHIR profiles. We used a view over the FHIR QI-Core profiles that hides FHIR-specific implementation details, notably FHIR extensions (e.g., birthtime). However, QI-Core profiles cannot shield implementers from other implementation challenges (e.g., FHIR cross-resource inconsistencies) because the QI-Core logical view is a one-to-one view directly based off of a specific version of FHIR. Moreover, because of the evolving nature of FHIR, as well as differences in how FHIR is implemented in EHRs, coupling a SMART on FHIR application with a specific FHIR version or implementation could be costly and potentially not scalable. Just during the course of this project, multiple versions of FHIR were released, with sometimes substantial changes occurring between releases. Use of a more consistent and FHIR-version-independent logical model could mitigate some of these challenges.
Results

Overview of Approaches Taken to Manage Functionality Versus Portability

We applied the following approaches for adapting the app for cross-institutional dissemination with the goal of balancing the functionality of the Bili App with its portability. We aimed to adequately meet the needs of clinician users while also maximizing the Bili App potential reach and impact. As described in detail below, the approaches involve (1) initial implementation based on meeting clinician needs, with custom FHIR APIs implemented as needed, (2) gathering feedback from current users through surveys and interviews about their valuation of specific app features, (3) enabling the application to provide differential features depending on available EHR FHIR capabilities, (4) progressive replacement of custom interfaces with native EHR FHIR interfaces as they became available, and (5) using a canonical logical data model that can be mapped to different FHIR versions and profiles.

Initial Implementation Based on Clinician Needs

We started by developing a fully-featured app based on user needs. To support this application, several custom FHIR APIs were developed to obtain required information not supported by the native EHR FHIR interfaces at the time, including for the patient’s gestational age, the mother’s laboratory data, the patient’s discharge time, and the timing of outpatient phototherapy orders. We developed these required FHIR APIs, leveraging and “wrapping” non-FHIR APIs natively supported by the EHR where possible. These custom FHIR APIs were made available to the Bili App alongside the native EHR FHIR APIs through an intermediate proxy.

Following initial deployment, additional improvements were made based on user feedback. Figure 1 shows the current version of the Bili App for a test patient. Numbers have been added to the screenshot corresponding to the key user interface components enumerated in the Methods.

![Figure 1. Screenshot of Neonatal Bilirubin Application for a Synthetic Patient](image)

The Bili App currently requires 15 data elements which are retrieved from the EHR as FHIR resources through native and custom APIs. Data elements and associated dissemination concerns are summarized in Table 1. In other health systems, different coding terminologies might be used to record the same data elements, or some elements might be unavailable.
### Table 1. Bili App Data Elements and Dissemination Concerns

<table>
<thead>
<tr>
<th>Data Element</th>
<th>FHIR Resource</th>
<th>Example of Query Parameters</th>
<th>API</th>
<th>Dissemination Concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth time</td>
<td>Patient</td>
<td>Patient/[id]</td>
<td>Native, custom</td>
<td>An Epic non-FHIR API is used to get the birth time and add it as an extension to the native FHIR resource</td>
</tr>
<tr>
<td>Hospitalization</td>
<td>Encounter</td>
<td>Encounter?type=</td>
<td>Native</td>
<td>Was not available in Epic 2017; now available</td>
</tr>
<tr>
<td>Albumin test</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Native</td>
<td>Mapping may be required*</td>
</tr>
<tr>
<td>Blood type</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Native</td>
<td>Mapping may be required*</td>
</tr>
<tr>
<td>Direct bilirubin test</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Native</td>
<td>Mapping may be required*</td>
</tr>
<tr>
<td>Direct Coombs test</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Native</td>
<td>Mapping may be required*</td>
</tr>
<tr>
<td>Gestational age</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Custom</td>
<td>Was not available in Epic 2017</td>
</tr>
<tr>
<td>Indirect Coombs test</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Native</td>
<td>Mapping may be required*</td>
</tr>
<tr>
<td>Total bilirubin test</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Native</td>
<td>Mapping may be required*</td>
</tr>
<tr>
<td>Transcutaneous bilirubin test</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Native</td>
<td>Mapping may be required*</td>
</tr>
<tr>
<td>Inpatient phototherapy</td>
<td>Procedure</td>
<td>Procedure?code=</td>
<td>Custom</td>
<td>Initially used custom FHIR service; can be replaced with native EHR Observation interface through LOINC mapping</td>
</tr>
<tr>
<td>procedure</td>
<td>Request</td>
<td>ProcedureRequest?code=</td>
<td>Custom</td>
<td>Requires custom FHIR service</td>
</tr>
<tr>
<td>Natural mother</td>
<td>Related Person</td>
<td>RelatedPerson?code=</td>
<td>Custom</td>
<td>Requires custom FHIR service</td>
</tr>
<tr>
<td>Mother’s blood type</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Custom</td>
<td>Requires custom FHIR service</td>
</tr>
<tr>
<td>Mother’s indirect Coombs test</td>
<td>Observation</td>
<td>Observation?code=</td>
<td>Custom</td>
<td>Requires custom FHIR service</td>
</tr>
</tbody>
</table>

*Mapping of local codes to LOINC may be required in EHR configuration

---

**Gathering Feedback from App Users**

User assessment of the value of app features is summarized in Figure 2. The numbers in the figure correspond to the feature labels included in the screenshot in Figure 1. Surveying for valuable features clearly showed that some features are perceived as more important than others. Two features were perceived as highly important by all participants: bilirubin results and phototherapy thresholds. Most of the users stated that they would still use the app if any one of the 8 difficult-to-disseminate features was omitted. We found that the hyperbilirubinemia risk tab and table of albumin measurements had medium to low value for most users. Auto-population of gestational age was valued highly by most users.
Figure 2. Participant Survey Responses

Interviews allowed us to get a deeper understanding of why users found some features less valuable than others. For example, users noted:

"I don’t use this tab at all <pointed to Hyperbilirubinemia Risk tab>. I know other people do on the inpatient side, but I’m not very familiar with it and haven’t used it to change my decision one way or another. Kind of redundant with this <main tab>..."

"We hardly ever draw albumin on babies. I feel like maybe for premies we might be looking at that, but we hardly ever have it on record."

In the interviews, providers tended to report what should be added, not what could be omitted. This is consistent with the psychological literature that users do not like to give away what they already have. Many users asked to add the rate of rise. For example, one user noted the following:

"It would be wonderful if it could automatically calculate rate of rise (the calculation is not hard or time consuming, but an automated value would be helpful for busy clinic days)."

Enabling Differential Features Based on EHR FHIR Capabilities

We are refining the system to be able to adapt to differing levels of data availability (Figure 3). For example, when a mother’s laboratory data cannot be automatically pulled through the FHIR interface, instead of leaving those data elements blank in the app, the app could be configured to not include a placeholder for those data elements in the first place. To an extent, the application is already configured to account for differences in FHIR data availability. For example, for patients who are not born at our healthcare system but are instead transferred in or simply followed up in the outpatient setting, the app prompts the user to enter the missing patient birthtime rather than pulling it from the EHR. We are planning to extend this type of differential functionality based on known and anticipated differences in FHIR capabilities across EHR systems. As an extreme case, we are also considering developing a stand-alone version of the app that can function without the provision of any EHR-provided data.
Progressive Replacement of Custom FHIR APIs with Native EHR FHIR APIs

During the course of the project, several FHIR APIs were introduced by the EHR vendor, which allowed for replacement of custom FHIR APIs. Data that could now be supported natively through use of the EHR vendor’s FHIR API include the patient’s gestational age and the date and time of the patient’s discharge from the hospital. In addition, we learned that inpatient phototherapy administration times could be pulled through a native EHR FHIR Observation API following configuration in the EHR to map nursing flowsheet data to LOINC. Thus, whereas a custom FHIR Procedure API was originally used to identify when phototherapy was administered in the inpatient setting, this information can now be retrieved using the native EHR FHIR interface for Observations.

Use of a Canonical Logical Data Model

To mitigate dissemination challenges associated with FHIR versioning, we propose using a canonical logical data model which allows the core SMART on FHIR application to be written against one data model rather than many data models, thus decoupling application logic from the FHIR versions used in messages. Currently, one of the authors (CN), who is a co-chair of the HL7 Clinical Information Modeling Initiative (CIMI) Work Group, is developing such a logical data model known as the HL7 Quality Improvement and Clinical Knowledge (QUICK) logical model. The QUICK logical model is based on multiple standards and models including HL7 FHIR,\(^1\) HL7 Virtual Medical Record (vMR),\(^19\) the National Quality Forum Quality Data Model (QDM)\(^20\), HL7 US Core FHIR profiles,\(^12\) and HL7 QI-Core FHIR profiles.\(^16\) The scope of the QUICK logical model consists of classes and attributes currently identified as needed for CDS and quality measurement. The QUICK logical model is under development at HL7 and is expected to be balloted as part of the HL7 January 2020 submission cycle.

Discussion

Summary of Findings

While SMART on FHIR provides a promising potential approach for widely disseminating innovative extensions to the EHR, the still evolving nature of EHR vendors’ support for FHIR poses challenges to such widespread dissemination. Specifically, in cases where needed FHIR APIs are not broadly supported across EHR vendors, a SMART on FHIR app developer must consider tradeoffs between functionality and ease of dissemination. In this manuscript, a neonatal bilirubin management SMART on FHIR application was used as a case study of how this...
tradeoff can be appropriately managed. The approach taken involved the following methods. First, we started by ensuring that user requirements are met in an initial implementation for production clinical use, with custom FHIR service being developed in coordination with the EHR vendor where needed. Second, surveys and interviews were conducted with users of this fully-featured application, with a goal of identifying those features that provide the least perceived value to end users, as those features could potentially be omitted when externally disseminating the tool if they require FHIR interfaces that may not be generally available. Third, we are enhancing the application so that its feature set can adapt to differing levels of data availability. Fourth, custom FHIR APIs were replaced with EHR vendor-provided interfaces as those interfaces became available, so that the need for developing or sharing custom FHIR APIs is reduced. Finally, we are moving towards the use of a canonical logical data model, so that the core logic can remain unchanged while enabling interaction with different FHIR versions as well as profiles supported by EHR vendors. While we have not yet empirically validated that these approaches will enable functionality to be optimized while still supporting widespread dissemination, we believe that these complementary approaches will enable us to achieve these goals moving forward.

Strengths and Limitations

One important strength of this study is that we used several complementary approaches spanning qualitative methods (interviews and surveys) as well as deeply technical approaches (e.g., developing custom FHIR APIs directly into EHRs and using a canonical logical data model). Thus, we were able to expand the potential approaches available to us. For example, if we had not possessed the capability of building custom FHIR APIs directly into the EHR for production clinical use, it is uncertain whether we could have developed a SMART on FHIR application for neonatal bilirubin management that would have adequately met the needs of our clinicians. Indeed, we had started by proposing clinical use of the SMART on FHIR demonstration application for this use case that had been available for some time on the SMART on FHIR app gallery. Despite the fact that this application had been successfully demonstrated across multiple EHR platforms, our clinicians felt that introducing this application into our clinical environment would provide little value without the enhancements we ultimately implemented. As a second strength, this study was conducted by a project team that included national experts on FHIR and its evolution in the standards community (KK and CN). One of the authors (KK) also serves on the Board of Directors of HL7 (the standards development organization that specifies FHIR) and the U.S. Health IT Advisory Committee. As such, the project team is well-positioned to understand the current state and future direction of the FHIR standard and its adoption by the health IT community and is actively contributing to its evolution through our implementation experience. As a final strength, our approach can be applied to a wide range of clinical use cases, as it allows new FHIR interfaces to be developed and used where needed. This approach does not rely solely on FHIR capabilities that are universally supported across EHR vendors, as our experience has indicated that many actual clinical needs cannot currently be adequately met using just these FHIR interfaces.

An important limitation of the study is that we are still in the process of disseminating the SMART on FHIR Bili App. Thus, we do not yet have empirical evidence that the approach used will be successful in balancing functionality with portability. A second limitation is that this approach involves making enhancements to the FHIR interfaces natively provided by the EHR vendor. Such enhancements could become a barrier to dissemination due to the cost of development and integration. For organizations that do not possess this capability, or whose EHR vendors do not permit such extensions, the only option remaining would be to advocate for the needed capabilities to be implemented by the EHR vendor. Finally, a third limitation of this study is that it is technically more complex to implement than an approach focused on using only those FHIR capabilities natively provided across EHR vendors. While this may be functionally adequate in some use cases, in other cases this would significantly hamper the functionality that can be achieved and the degree to which duplicate manual data entry can be avoided.

Lessons Learned and Recommendations

Through this study, several insights were gained on the process of balancing functionality and portability. First, while improving, existing FHIR capabilities provided by EHR providers are often inadequate for meeting the needs of clinical users. Second, the degree to which EHR vendors support and empower their clients to enhance FHIR interfaces is variable. Third, while most EHR vendors provide support for US Core FHIR profiles, these profiles still provide optionality for EHR vendor implementation, leading to vendor differences in implementation. Fourth, mapping of local codes to standard codes is still oftentimes required. Finally, while promising, enabling widespread dissemination of functionally rich SMART on FHIR applications can still present significant challenges.
Based on our study findings and lessons learned, we recommend that stakeholders in the healthcare community cooperate in identifying their most important needs with regard to FHIR and advocate for their universal support in the US Core FHIR profiles and EHR vendor implementations of those profiles. There is substantial ongoing work in this area, for example by the U.S. Core Data for Interoperability and Interoperability Standards Priorities Task Forces of the U.S. Health IT Advisory Committee. Moreover, given the limitations of US Core FHIR Profiles in addressing end user requirements, we (CN and KK) are currently proposing FHIR Implementation Guides that extend beyond US Core FHIR Profiles for adoption by the Healthcare Services Platform Consortium (HSPC), Clinical Information Interoperability Council (CIIC) and HL7 as part of the Reimagine EHR initiative.

Interested stakeholders are encouraged to engage in these forums, as well as directly through HL7 and other venues to advocate for their priorities. Second, we recommend that SMART on FHIR app developers explicitly consider the trade-offs between functionality and portability, as well as to use one or more of the approaches described in this study for managing those trade-offs. Finally, we recommend that end-users be engaged throughout the process to ensure that the approach taken adequately meets their clinical and workflow needs.

Future Directions

In the future, approaches to measuring the cost and utility of providing desired features that extend beyond US Core FHIR profiles should be explored. Such measurement could allow formal evaluation of the extent to which the benefits of adding such features outweigh the costs. Expanding such cost-benefit analysis on the national level could help identify priorities for the development and implementation of standards including the US Core FHIR profiles.

Moving forward, we are continuing to develop new FHIR capabilities in the EHR, including SMART on FHIR applications and CDS Hooks decision support services. Other applications that we have developed or are currently developing for clinical use include a population health management system for individuals at risk of early onset familial cancers, a surgical referral dashboard, a procedure capacity management application, a lung cancer screening shared decision making app, a diabetes treatment outcome prediction app developed in collaboration with Hitachi, opioid-related decision support tools developed with CDC and ONC support, and EHR-integrated medical calculators developed in collaboration with MDCalc. We believe in the potential for SMART on FHIR applications to improve patient care and reduce physician burnout. As one physician noted when interviewed regarding the Bili App:

“We need to be thinking way beyond the bilirubin app. The sky is the limit.”

Conclusion

As leaders in SMART on FHIR application development, we share here our journey in preparing one of our first SMART on FHIR applications intended for wide dissemination. We demonstrated the feasibility of satisfying user requirements through developing custom FHIR APIs not supported natively by our EHR. Unfortunately, this approach would require a significant investment of resources on the part of adopting organizations and thus poses a barrier to dissemination and adoption of the application. To address this challenge, we recommend that such FHIR services be considered as a starting point for the development of future FHIR implementation guides that extend or build upon the US Core FHIR profiles. The implementation of such standards-based APIs by EHRs would reduce the barriers to dissemination. Until such a time, a holistic approach that combines both technical and user-centered qualitative methods is recommended for optimizing functionality versus portability for SMART on FHIR applications.

References

4. Mandl KD, Mandel JC, Kohane IS. Driving Innovation in Health Systems through an Apps-Based


Interpretation of machine learning predictions for patient outcomes in electronic health records

William La Cava, Ph.D.¹, Christopher Bauer, Ph.D.², Jason H. Moore, Ph.D.¹, Sarah A Pendergrass, Ph.D.²

¹University of Pennsylvania, Philadelphia, PA, USA; ²Biomedical and Translational Informatics Institute/Geisinger, Danville, PA, USA

Abstract

Electronic health records are an increasingly important resource for understanding the interactions between patient health, environment, and clinical decisions. In this paper we report an empirical study of predictive modeling of seven patient outcomes using three state-of-the-art machine learning methods. Our primary goal is to validate the models by interpreting the importance of predictors in the final models. Central to interpretation is the use of feature importance scores, which vary depending on the underlying methodology. In order to assess feature importance, we compared univariate statistical tests, information-theoretic measures, permutation testing, and normalized coefficients from multivariate logistic regression models. In general we found poor correlation between methods in their assessment of feature importance, even when their performance is comparable and relatively good. However, permutation tests applied to random forest and gradient boosting models showed the most agreement, and the importance scores matched the clinical interpretation most frequently.

Introduction

Electronic health record (EHR) adoption grew from 9.4% to 83.8% in hospitals across the United States over the last decade, mostly due to incentives provided by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009¹. Thus for the foreseeable future, EHR data will be one of the most comprehensive and promising resources for understanding the interactions between medical care and patient health and outcomes. EHR data integrate disparate information about patients and their health history, including patient demographic data, diagnoses, medication prescriptions, clinical lab measures, physician’s notes, and radiological imaging data². An increasing number of health care systems are integrating genetic data as well³, which allows data analysis to span biology and medical informatics.

EHR data can support a number of knowledge discovery tasks such as predictive modeling, identifying disease co-morbidity, and identifying patient sub-types. Besides its use in diagnosis, this knowledge can be leveraged to generate new hypotheses or improve clinical trial design². This paper focuses on the first task: predictive modeling of patient outcomes, which can impact care in several ways. First, models that predict patient outcomes can be used at the point of care for assisting in clinical decision making. Second, the models can be used to identify trends in undesirable patient outcomes and support new clinical procedures that amend those trends. Third, predictive models can point to underlying factors that affect a patient subgroup’s response to a treatment, thereby improving scientific understanding of human health.

Despite the potential for EHR data, current statistical and machine learning (ML) methods are limited in their capacity to learn from these data for a variety of reasons. Many commonly used methods do not natively handle mixed data types or longitudinal data collected at non-uniform intervals. More fundamentally, it is not known a priori what underlying structure a given collection of EHR data might contain, and thus the best choice of ML method is non-trivial. This uncertainty has pushed the research field towards black-box, high-capacity methods embodied by deep learning. The common deep learning approach to EHR analysis is to develop unsupervised auto-encoders¹,4,5 that perform dimensionality reduction. Although potentially useful, the resulting data representations may not improve predictive models of patient outcomes. Another deep learning approach is to train ensembles of large recurrent neural networks, an approach shown to yield slight improvements over baseline logistic regression models⁶. In either case, the crucial task is to identify explanatory features of models² and interpret them. This task has been noted as a weakness of state-of-the-art approaches using deep learning⁷.

In order to address interpretability, we focused our analysis in this paper on the interpretation of various ML models
Table 1: Data used for predictive modeling.

<table>
<thead>
<tr>
<th>Item</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease (ICD-9 code)</td>
<td>Liver disease (571.8), Alzheimer’s disease (331.0), Kidney disease (585.9), Diabetes with renal manifestations (250.40), Sleep apnea (327.23), Diabetes (250.00), Esophageal reflux (530.81)</td>
</tr>
<tr>
<td>Demographics</td>
<td>age, sex, race, ethnicity</td>
</tr>
</tbody>
</table>

for the task of disease prediction. We trained 3 state-of-the-art ML methods to predict 7 patient diagnoses with varying prediction horizons. For each ML model that was developed, we compared model-specific and model-agnostic feature importance scores, including coefficient importance, Gini importance, univariate effect size, and permutation importance. We then conducted a correlation and interpretability analysis for 7 diseases to determine 1) how well the important features comport with reality, 2) how well different measures of importance agree, and 3) how well ML models agree, taking into account their predictive performance. The main finding of our analysis is that permutation importance, a model-agnostic method, produces the most clinically relevant interpretations, as long as the underlying model produces good predictions. Among models with high predictive performance on test sets, permutation importance scores were highly correlated, and interpretable in the sense that they matched clinical understandings of these diseases.

Methods

In this section we describe the data resource and the preliminary data processing used to formulate the prediction task. We then describe the ML approaches and feature importance measures used in this paper, along with their strengths and weaknesses. Code to reproduce the analysis in this paper is available online.

Geisinger Health Records

We conducted our analysis on a set of data from 899,128 patients from the Geisinger Health System, collected between 1996-2015. For each patient we collected their available lab measures, demographic information and diagnostic codes, summarized in Table 1. We split the laboratory measures into two groups: common measures, for which the missing rate was less than 46.5% across patients, and rare measures. This threshold was chosen via a sample size analysis by Beaulieu-Jones et. al.8. For the common measures, the median lab values were used as predictors. Missing values for these measures were imputed using softImpute9. The choice of imputation algorithm was made based on a previous study of imputation methods, conducted on the same patient population8.

We determined disease status using ICD-9 diagnosis codes. These diagnosis codes represent 80% of the total diagnosis codes in the EHR, and all codes between 1996 and 2015. ICD-10 codes were available for patients with recent encounters; however, since these codes are non-trivial to integrate, we left this task to a future study. In order to

---

1http://github.com/EpistasisLab/interpret_ehr
improve our confidence in ICD-9 codes as a proxy for diagnosis, we required that cases had 3 instances of these codes in their medical record. Controls were conversely selected from those patients that had no instances of the corresponding code in their record.

We constructed the learning task as follows: given a patient’s record some length of time before their diagnosis, predict their outcome status. We constructed predictive models with a prediction horizon of one day, six months, and one year prior to diagnosis to assess the sensitivity of our predictions to recent information about the patient. Since some patients did not have up to one year or six months’ worth of medical records, the cohort for each studied disease shrank slightly as the horizon increased. However, this effect was minimal as shown in Fig. 1.

**Figure 1:** Number of cases for each disease and prediction horizon.

**Machine Learning Methods**

In order to classify a patient’s disease status, we build a classification model $\hat{y}(x)$ trained on a labelled set of training examples, $\{y_i, x_i\}_{i=1}^N$. Each of the $N$ examples represents a patient, where $x \in \mathbb{R}^d$ is a $d$-dimensional vector of predictors (from Table 1) and $y \in \{0, 1\}$ is the patient’s outcome, encoded as 1 if the patient is diagnosed and 0 otherwise. We use $X$ to refer to a matrix of predictors/features, with $N$ rows and $d$ columns.

We analyzed three ML methods for predicting patient outcomes: penalized logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost). We chose to use LR and RF due to their pervasiveness and accessibility to researchers. XGBoost was chosen as well due to its strong performance in many recent competitions, and subsequent adoption as the out-of-the-box classifier of choice$^{10}$. For each of these methods, we briefly summarize their methodology below and discuss how one may interpret the models that result.

**Penalized Logistic Regression**

Logistic regression trains a linear model on the log-odds ratio of the outcome being positive, i.e.

$$\log \left( \frac{Pr(y_i = 1|x_i)}{Pr(y_i = 0|x_i)} \right) = \beta^T x_i$$

where $\beta = [\beta_1, \ldots, \beta_d]$ are coefficients associated with each predictor. In our analysis, we assumed the predictors are standardized to unit variance and mean-centered so that the intercept is zero.

In standard logistic regression, the coefficients are chosen to maximize the log-likelihood of the observations. Penalized regression applies an additional penalty term that is proportional to the magnitudes of the coefficients, as:

$$\max_\beta \left\{ \sum_{i=1}^N [y_i(\beta^T x_i) - \log(1 + \exp(\beta^T x_i))] - \lambda \sum_{j=1}^d ||\beta_j|| \right\}$$

(2)

The $\lambda$ term determines the strength of regularization; a larger $\lambda$ forces the coefficients to be closer to zero. The norm $|| \cdot ||$ may be the L1 norm (lasso$^{11}$) or L2 norm (ridge regression$^{12}$). The coefficients in Eqn.1 can be interpreted
as the change in the log-odds of the outcome per unit change in the corresponding predictor, correcting for all other predictors. As a result, LR is one of the most popular methods for interpretable modeling.

In addition to the log-odds interpretation, it is common to use the coefficient magnitudes ($|\beta|$) as an estimate of the importance of each predictor. With standardized predictors this seems reasonable, since the factors with the largest effect sizes have the most influence on the log-odds ratio of the prediction. However, as others have pointed out, this interpretation has serious caveats\cite{13, 14}. Most notably, when model assumptions are violated, the coefficient values will be incorrect. In the presence of multicollinearity, for example, the coefficients may compensate for correlations in $x_i$ when maximizing the objective function (Eqn. 2). In this scenario, it is no longer valid to interpret their effect on the log-odds ratio of the outcome independently, since the predictors are dependent on each other.

Many techniques to address multicollinearity of regression models have been proposed\cite{13, 15}. For example, one can drop features with high variance inflation factors, add interaction terms, or conduct feature selection to remove correlated variables. Penalized regression tends to improve estimates of the coefficients in cases of multicollinearity since the penalization term improves the condition of $X$ from which the estimates are derived\cite{12}. Nevertheless, the success of these techniques must be validated empirically. In our analysis, we considered lasso and ridge regression with varying degrees of penalization ($\lambda$), using both the traditional coefficient magnitude interpretations as well as a model-agnostic permutation test that is described below.

**Random Forest**  Random forest is an ensemble ML model that trains several decision trees using a combination of bootstrap aggregating (a.k.a. bagging) and random feature selection\cite{16}. The final model output is determined by a majority vote of the outputs of the individual trees. One of the attractive features of RF is the ability to estimate the importance of each features in the trained model. This feature importance is known as the Gini Importance.

Decision trees (the basic ML models comprising the ensemble) use a heuristic to determine which feature to split on while recursively constructing the model; in our case this heuristic is the Gini criterion. By storing these heuristic measures at each node, the importance of each feature can be estimated quickly from the model. The Gini importance of each feature is estimated by measuring the mean decrease in the heuristic that is brought about by splitting on that feature in any place within the forest. This score is normalized across the forest and across features so that all Gini importance scores sum to 1.

Several authors have pointed out issues with this method of determining feature importance\cite{17–20}. For one, features with a higher number of split points (e.g. continuous features) are more likely to have a high importance purely due to the number of splits. This is due to sampling bias: an optimal split chosen among more candidate points is more likely to reduce the Gini criterion purely by chance\cite{17, 20}. In addition, because feature importance is defined relative to the training data, the bootstrap sampling approach utilized by RF can introduce a bias: for a given training instance, only certain variables and/or levels of variables will be competing when the optimal split point is chosen during tree construction. This has led to the observation in genetics studies that Gini importance prefers SNPs with larger minor allele frequencies\cite{21}.

**XGBoost**  XGBoost is an ensemble ML method based on gradient boosting of individual decision trees. Rather than simultaneously training a forest of trees like RF, gradient boosting creates an ensemble by iteratively training decision trees on weighted training samples, where the weights are updated each iteration to reflect the residual error of the current ensemble. XGBoost uses a form of regularized gradient boosting proposed by Friedman et. al.\cite{22} and includes additional optimizations that have led to its prominence among the leading entries to several ML competitions\cite{10}. By default, XGBoost also uses Gini importance as an internal feature importance score.

**Other feature importance measures**

In addition to the internal feature importance measures from constructed models, we considered two other approaches: a univariate regression score and permutation importance.
**Univariate Score**  For this score we simply calculate the magnitude of the marginal effect of each standardized predictor in \( x \) by fitting univariate LR models and filtering out insignificant coefficients according to a \( p \)-value threshold of 0.05. This is the simplest feature importance measure tested here, and unsurprisingly has strong assumptions, namely that a predictor’s importance is independent of all other factors. It is also important to note that significant predictors may not make useful predictions, as discussed in other work\(^2\).

**Permutation Importance**  We use a model-agnostic permutation importance score first proposed by Breiman et al.\(^6\) to estimate the importance of the features in the trained models. Permutation importance is defined as the mean decrease in accuracy of the trained model when each feature is permuted. We calculate the permutation importance of predictor \( x_j \in x \) by the following steps:

1. Create a permuted test set \( \{y_i, x'_i\}_{i=1}^{N_t} \) in which \( x_j \in x \) is randomly shuffled. \( N_t \) is the number of test samples.
2. Generate predictions on the normal test set, \( \hat{y}(x) \), and permuted predictions, \( \hat{y}(x') \)
3. The permutation importance (PI) is the mean decrease in accuracy due to the perturbed feature, i.e.:

\[
PI(x_j) = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{I}[y_i = \hat{y}(x_i)] - \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{I}[y_i = \hat{y}(x'_i)]
\]

Permutation importance has a few desirable properties. By using a randomly shuffled predictor as the permutation, permutation importance score compares the importance of each feature to an identically distributed predictor, thereby reducing potential bias. Furthermore, this score is produced on test data which renders a more accurate portrait of how the model behaves with new data. Permutation importance also allows us to make apples-to-apples comparisons of the importance of different ML models trained on the same data.

Downsides of permutation testing include its complexity and its inability to handle variable interactions. Permutation importance scores require generating predictions on the test set twice for each predictor, which may be computationally intractable for larger feature spaces. The permutation scores also do not take into account that predictors may naturally vary together. This can cause misleading interpretations for certain models. As an example, consider a model with two correlated features. Although the model may assign a weight of zero to one of the features if they always change together, a permutation test would indicate that one of the correlated features is unimportant. For the model this is true but the context is important. Therefore, the interpretation of permutation importance scores must take into account whether predictors are expected to change independently.

Permutation importance has been extended to address correlated variables\(^1\) and also can be extended to incorporate statistical tests\(^5\). Many other model-agnostic measures of feature importance exist, such as Relief and its variants\(^2\). However, Relief scales with the number of samples squared, making it intractable for tens to hundreds of thousands of patients.

**Experimental Setup**

For each outcome, we created an evaluation cohort of cases and controls. We used all cases, and sub-sampled the controls to match the sex and age quartile of each case. This results in datasets twice the size of the samples in Fig. 1 with 147 total predictors. To evaluate each ML method, we conducted 10 repeat trials on random shuffles of the data. For each trial, the data was split 50/50 into training and test sets. We conducted hyper-parameter tuning via 10-fold cross validation on the training set for each method. We then evaluated the tuned model on the test set and reported the area under the receiver operator characteristic curve (AUROC).

The internal feature importance measure of the tuned method applied to the entire training set was stored, in addition to permutation importance scores for the model on the test set. For each patient outcome, we examined the important features and assessed their interpretability. We also looked at the correlation between feature importance scores in a pairwise fashion to determine agreement between the models.
**Results**

The results for each disease, grouped by ML method and prediction horizon, is shown in Fig. 2. We find across diseases that XGBoost (XGB) performs the best, followed closely by RF. LR under-performs these tree-based ensemble methods, especially for predicting Alzheimer’s disease and esophageal reflux. The differences in performance between the methods on all diseases at each prediction horizon are significant ($p < 9.5 \times 10^{-12}$) according to pairwise Wilcoxon rank-sum tests with Bonferroni correction.

We find reasonably high AUROC performance across diseases. With a one year prediction horizon, the best model (XGBoost) achieved median AUROC values of 0.85 for Alzheimer’s disease, 0.91 for Diabetes, 0.96 for Diabetes with renal manifestations, 0.85 for esophageal reflux, 0.91 for kidney disease, 0.88 for liver disease, and 0.87 for sleep apnea. We find that predictive ability diminishes as the prediction horizon increases, as expected. The exception is the predictions for esophageal reflux that trend higher for the ensemble methods with longer prediction horizons.

We find that, across diseases, the Gini importance scores for XGBoost do not align very well with the expected importance scores. An example of this is shown in the scatter plots in Fig. 3, which shows XGBoost and RF importance measures using both Gini importance (top) and permutation importance (bottom) for the diabetes models. From left to right, the prediction horizon increases. We expected the presence of a hemoglobin A1c lab to be very predictive of outcome one visit prior to diagnosis, and for other risk indicators such as high average glucose levels, high BMI, and high triglycerides to be more predictive at six months to one year. We see this behavior with the RF importance scores to an extent, but observe less intelligible Gini importance scores from the XGBoost model. The XGBoost Gini importance scores suggest the anion gap measure (a potential sign of diabetic ketoacidosis) and age (a universal risk factor) are the most important. However, the permutation importance measure applied to both models (bottom) generated good agreement with expected predictors, and also shows that the two models actually agree to a large extent about which factors are important.

XGBoost tends to over-estimate the importance of age on nearly every outcome, likely due to the bias of the Gini importance measure discussed earlier. Age is a continuous variable in our analysis (it is calculated using the visit date), so despite being corrected for via quartile-matching among the controls, there appears to be enough variability.
Figure 3: Pairwise comparisons of feature importances for predicting diabetes. The prediction horizon increases to the right. Top: Gini importance scores for Random Forests and XGBoost. Bottom: Permutation importance for the same models. The permutation importance scores show better agreement between the models and produce more intuitive clinical interpretations.

Across outcomes, RF permutation and XGBoost permutation scores are very correlated (Pearson’s $R^2=0.95$) as shown by the pairwise comparisons in Fig. 4. The Gini importance measures for these models show somewhat lower correlation with each other ($R^2=0.85$), and the correlation with LR coefficient magnitudes (LR-importance in Fig. 4) is lower still ($R^2=0.37$ with RF and $R^2=0.34$ with XGBoost). In addition, the permutation importance of the LR models is uncorrelated with all other importance measures ($R^2 \leq 0.55$) aside from the LR coefficient magnitudes ($R^2=0.81$). This may be due to in part to the lower performance of these models for prediction (see Fig. 2) and due to the conflicting behavior of LR and permutation importance in the presence of collinearity, discussed in the methods section. Finally, the univariate importance score produce the importance scores that agree least with the other approaches, suggesting that a univariate analysis is insufficient for determining the important factors of these diseases.

We summarize the important features for each disease prediction using RF permutation importance in Fig. 5, with one visit predictions on the left and one year prediction models on the right. Diseases and their most important predictors are clustered. The predictions from one visit prior to diagnosis are the most interpretable; diabetes and diabetes with renal manifestations cluster based on the HbA1c test, with microalbumin indicating kidney complications for the latter diagnosis. Liver disease and sleep apnea cluster based on the importance of BMI. Alanine aminotransferase is the next most important predictor of liver disease, as expected. Kidney disease predictions depend on creatinine, urea nitrogen, age and glucose measurements, which are all expected. Alzheimer’s disease predictions depend on age in addition to cobalamin (vitamin B-12) measures. Cobalamin tests are typically ordered to rule out symptoms of a cobalamin deficiency (e.g. memory loss) prior to Alzheimer’s diagnosis. The most difficult interpretation is for esophageal
**Discussion**

EHR data can be used to support the development of predictive models that may assist clinicians at the point of care, as well as provide insight into factors driving outcomes. In this paper, we evaluated the ability of three state-of-the-art ML approaches to produce interpretable predictive models for 7 patient outcomes, considering prediction horizons of up to one year from diagnosis. We focus on methods for assessing feature importance of the constructed models, including model-specific and model-agnostic approaches. We find that gradient boosting (XGBoost) generates the most accurate predictions across outcomes, but that its internal measure of feature importance (Gini importance) is insufficient for a reliable clinical interpretation of the model. Applying the model-agnostic permutation importance score to the resultant models fixes this shortcoming and results in models with sensible interpretations. Permutation
importances match clinical intuition and agree between the XGBoost and RF models.

We plan to address the following shortcomings of this analysis in future work. First we would like to include additional predictive factors in our analysis, including smoking status, socio-economic measures, and disease comorbidities. We plan to incorporate ICD-10 disease codes from 2015-present available in Geisinger’s health records for assessing diagnoses. We also plan to expand the number of diseases studied and the prediction horizon to characterize the point at which models are no longer predictive of outcome. Additionally, we plan to validate the constructed models on a separate patient population to determine how generalizable these predictive models are given the idiosyncrasies of individual health care systems.

Acknowledgments

This work was supported by NIH grants K99LM012926, LM010098, AI116794, and TR001878, and a PA-CURE grant from the Pennsylvania Department of Health. This project is funded, in part, under a grant with the Pennsylvania Department of Health (#SAP 4100070267). The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

References


Interpretation of ‘Omics dynamics in a single subject using local estimates of dispersion between two transcriptomes

Qike Li1,2,4,8, Samir Rachid Zaim1,2,4,§, Dillon Aberasturi1,2,4, Joanne Berghout1,3, Haiquan Li3,2,4, Francesca Vitali1,2, Colleen Kenost1,2,5, Helen Hao Zhang4,7,*, Yves A. Lussier1,6,*

1Center for Biomedical Informatics and Biostatistics(CB2), 2Department of Medicine, 3Center for Applied Genetics and Genomic Medicine, 4Graduate Interdisciplinary Program in Statistics, 5BIOS Institute, 6University of Arizona Cancer Center, 7Department of Mathematics, The University of Arizona, Tucson, AZ 85721, USA;
* Corresponding authors; § Contributed equally

Abstract

Calculating Differentially Expressed Genes (DEGs) from RNA-sequencing requires replicates to estimate gene-wise variability, a requirement that is at times financially or physiologically infeasible in clinics. By imposing restrictive transcriptome-wide assumptions limiting inferential opportunities of conventional methods (edgeR, NOIseq-sim, DESeq, DEGseq), comparing two conditions without replicates (TCWR) has been proposed, but not evaluated. Under TCWR conditions (e.g., unaffected tissue vs. tumor), differences of transformed expression of the proposed individualized DEG (iDEG) method follow a distribution calculated across a local partition of related transcripts at baseline expression; thereafter the probability of each DEG is estimated by empirical Bayes with local false discovery rate control using a two-group mixture model. In extensive simulation studies of TCWR methods, iDEG and NOIseq are more accurate at 5%<DEGs<20% (precision>90%, recall>75%, false_positive_rate<1%) and 30%<DEGs<40% (precision>recall>90%), respectively. The proposed iDEG method borrows localized distribution information from the same individual, a strategy that improves accuracy to compare transcriptomes in absence of replicates at low DEGs conditions.

http://www.lussiergroup.org/publications/iDEG

Keywords: transcriptome, mRNA expression, RNA-Seq, differentially expressed genes, single-subject, N-of-1, RNA-Seq, iDEG.

Introduction

Precision medicine aims to deliver “the right treatments, at the right time, to the right person” [Kaiser 1]. However, clinical research, medicine, and pharmacology need new tools to achieve that goal. The prevailing system of one-size-fits-all drug development has led to the ten top-grossing USA drugs being ineffective for more than 75% of users, and these patients typically cannot be identified until after therapeutic failure has occurred. The success of precision medicine hinges on identifying the precise aberrant mechanisms at play during an individual’s disease course to optimize treatment based on that individual’s biology.

Single-subject RNA sequencing (RNA-Seq) analysis considers one patient at a time, with the goal of revealing an individual’s altered transcriptomic mechanisms. Relative to traditional cohort-based analyses, a major challenge of single-subject RNA-Seq analysis is the estimation of gene expression variance which is required to identify differentially expressed genes (DEGs). In cohort-based methods, gene variance is calculated across a heterogenous set of samples, and the statistical methods employed leverage and rely on those replicates. However, they also emphasize consistent and average responses which may not accurately represent a single patient when the disease is heterogenous or stratified. Alternatively, the variance can be assessed between two conditions in one subject and three replicates. Yet, obtaining sufficient isogenic replicates for one subject to answer more precision questions poses a major difficulty due to (i) limited tissue availability, (ii) the risks associated with invasive tissue-sampling procedures, and (iii) general costs and inefficiencies with the current technology. Even though there is a great body of work for identifying DEGs in RNA-Seq data—even frameworks for N-of-1 studies either for a single analyte or by pooling gene products in pathways, to the best of our knowledge, no methods have been designed or validated at the gene level to determine the effect size and statistical significance of a single-subject, single RNA-Seq studies in two conditions without replicates (TCWR). Strategies to implement standard RNA-seq analysis methods for comparing TCWR have been proposed in the respective methods’ publications without comprehensive evaluation. Typically, these standard methods, usually requiring large cohorts, have been adapted to identify DEGs in TCWR by imposing restrictive transcriptome-wide distribution assumptions, thus limiting localized inferential opportunities.
Three critical obstacles hinder the analysis of single-subject TCWR studies. These include i) patient-level inferential capability in absence of biological replicates, ii) sensitivity to fold-change inflation in low-expression genes, and iii) rigid parametric data assumptions for variance estimation. To overcome the current technical limitations in analyzing RNA-Seq data, we propose a new method that borrows localized information across different genes from the same individual using a partitioned window to strategically bypass the requirement of replicates per condition: iDEG (individualized Differentially Expressed Genes). iDEG applies a localized variance-stabilizing transformation to estimate a gene’s distribution that borrows information from genes with similar baseline expression. While variance-stabilizing transformation has been previously used to identify DEGs across a large number of subjects or replicates, our approach differs from these since it has been developed to be applied directly on two paired transcriptomes from a single subject by computing the localized dispersion parameters in different windows of genes with similar expression at the baseline.

In this work, we evaluated the performances of iDEG and other four standard approaches applied to single-subject TCWR studies (edgeR, NOISeq-sim, DESeq, DEGseq). We also designed simulation studies under several conditions to stratify the range of applicability of our proposed strategy, which could eventually complement other RNA-seq analyses in TCWR studies. This study demonstrates the utility of variance-stabilizing transformations within subject in absence of replicates in two conditions, which is distinct from previous implementations of variance-stabilizing methods conducted across replicates or subjects.

Methods

The iDEG algorithm: iDEG

The iDEG algorithm (Figure 1) is an easy-to-implement, single-command function written in R with a computation speed of one second for identifying a subject’s DEGs on 8GB Ram computer. The subsequent sections expand on the main iDEG steps shown in Figure 1.

1. Normalize Library Sizes
2. Partition genes into windows
3. Obtain window-specific estimates
4. Apply VST function, h()
5. Compute Z-statistic
6. Estimate locfdr

Modeling read counts via a re-parameterized Negative Binomial (NB) distribution

We model read counts \( Y_{gd} \) as following a re-parameterized negative binomial distribution with mean \( \mu_{gd} \) and dispersion \( \delta_g \). Thus, \( Y_{gd} \sim \text{NB}(\mu_{gd}, \delta_g) \) with the following probability mass function, mean, and variance, respectively. Since for
any subject, both transcriptomes are sequenced separately, they are treated as conditionally independent, conditional on the subject*.  

\[
P(Y_{gd}|\mu_{gd}\delta_g) = \left(1 + \delta_g\mu_{gd}\right)^{\frac{1}{2}} f\left(Y_{gd}\right) \left(\frac{\delta_g\mu_{gd}}{1+\delta_g\mu_{gd}}\right)^{Y_{gd}} ; E(Y_{gd}) = \mu_{gd}, \ Var(Y_{gd}) = \mu_{gd} + \delta_g\mu_{gd}^2 \quad \text{(Eq. 1)}
\]

To identify DEGs from a pair of transcriptomes, we must test multiple hypotheses \(H_g; \mu_{g1} = \mu_{g2}; \ g = 1, \ldots, G\), where \(\mu_{g1}\) and \(\mu_{g2}\) are the theoretical mean expression levels for each gene “g” in sample 1 and sample 2, respectively. We define the DEG set by \(G = \{g: \mu_{g1} \neq \mu_{g2}; \ g = 1, \ldots, G\}\) and its set-theoretic complement of non-differentially expressed genes, or “null gene set” by \(G' = \{1, \ldots, G\}\setminus G\). In presence of replicates, each hypothesis can be tested with a two-sample comparison, using Welch’s t-test statistic: \(t_g = (\bar{Y}_{g1} - \bar{Y}_{g2}) / [(S^2_{g1}/N_{g1} + S^2_{g2}/N_{g2})^{1/2}]\) were \(\bar{Y}_{g1}\) and \(\bar{Y}_{g2}\), \(S^2_{g1}\) and \(S^2_{g2}\), and \(N_{g1}\) and \(N_{g2}\) are each groups’ respective sample mean, standard deviation, and size. However, when there is only one observation for \(Y_{g1}\) and one for \(Y_{g2}\), neither \(S^2_{g1}\) nor \(S^2_{g2}\) are computable. We thus propose iDEG: an algorithm that transforms \(Y_{g1}\) and \(Y_{g2}\), such that a simple function of the transformation allows for modeling all genes with the same distribution. This is done by pooling the genes together and estimating their common variance, hence bypassing the single-subject, single-replicate limitation.

**Normalize read counts with unequal library sizes (Figure 1, Panel 1)**

In practice, unequal DNA library sizes may exist; thus, the first step is to normalize library sizes if necessary. We use the quantile-adjusted conditional maximum likelihood (qCML) procedure by Robinson and Smyth *, for normalization, and subsequently, iDEG is applied.

**Partition genes into windows to estimate local mean and variance (Figure 1, Panel 2)**

Marioni et al. demonstrated the aptness of using expression means to estimate a gene’s variance*. Therefore, by extension, in iDEG we assume that genes of comparable expression levels are assumed to behave similarly (genes with similar means share similar variances). Thus, after normalization, the next step is to group genes into \(W\) non-overlapping windows of similar expression levels to approximate each window’s local mean and variance parameters. In the re-parameterized NB distribution, the variance of a given gene, \(g\), is a function of its mean, \(\mu_g\), and dispersion, \(\delta_g\). Thus, the genes are partitioned to obtain their local, window-specific parameters. We define the \(w^{th}\) window by the \((w-1)^{th}\) and \(w^{th}\) percentiles, for \(w = 1, \ldots, W\), \(G_w = \{g: (w-1)^{th} \text{ percentile of } Y_{g1} < Y_{gw} < w^{th} \text{ percentile of } Y_{g1}\}\). To provide robust parameter estimates, we recommend a large positive integer, \(W\), so that each window contains between 150 and 200 genes. However, the final predictions are not overtly sensitive to the choice of \(W\) (< 10% difference, data not shown – available upon request).

**Compute each window’s parameters (Figure 1, Panel 3)**

As seen in (Eq. 1), \(\text{Var}(Y_{gd}) = \mu_{gd} + \delta_g\mu_{gd}^2\). Therefore, when estimating variance locally, we are actually estimating the dispersion parameter \(\delta_g\) for each gene count. This local estimation provides a more numerically fair evaluation of genes as it allows for comparisons relative to their mean expression counts. Particularly, it enables a better estimation of dispersion and variance parameters for genes with extremely high or low expression counts, since these genes are grouped together into windows, and share window-level parameter estimates in order to over-inflate or deflate their variability by averaging it out across the entire transcriptome. This is done in effort to mitigate challenges with making DEG calls in lowly and highly expressed genes. This value is required for the variance-stabilizing transformation (VST) calculation (Eq. 3; Figure 1 Panel 4). However, variance cannot be estimated when only a single observation is available. For RNA-Seq data analysis, one common assumption is that the dispersion \(\delta\), is equal across samples 1 and 2, and that dispersion is a function \(q\) of the mean, \(\mu_g\). Thus, in the absence of replicates, we partition genes into small windows to estimate the functional mean-dispersion relationship, \(\delta = q(\mu)\), and hence the variance. We propose a two-step nonparametric procedure to obtain: (i) an initial estimate of \(\delta\) by pooling genes locally; and (ii) a refined estimate of \(\delta\), by estimating \(q(\mu)\) with a smooth curve-fitting technique. In this approach, all non-differentially

---

* A note on notation, since iDEG models each patient’s paired transcriptome individually, the subscripts for each subject are omitted since only a subject is handled at a time in any given calculation.
expressed (null) genes belonging to the same window \( G_w \) roughly have the same mean \( \mu_w^* \) and the dispersion value \( \delta_w^* \). Thus, \( Y_{gd} \sim NB(\mu_w^*, \delta_w^*) \) \( \forall g \in G_w, d = 1, 2 \) where \( \mu_w^* \) and \( \delta_w^* \) are the window-specific mean and dispersion values for null genes in \( G_w \), while \( d \) specifies if the count comes from sample 1 or 2.

The initial window estimates \( \mu_w^* \) and \( \sigma^2_w \) as \( \hat{\mu}_w = \text{Median}(Y_{gd}) \) \( \forall g \in G_w \) and \( \hat{\sigma}^2_w = \text{Median}(|Y_{gd} - \text{Median}(Y_{gd})|) \) \( \forall g \in G_w \), respectively. Since \( \text{Var}(Y_{gd}) = \sigma^2_{gd} = \mu_{gd} + \delta_{gd} \mu_{gd}^2 \), we estimate \( \delta_w^* \) with \( \hat{\delta}_w^* = (\sigma^2_w - \hat{\mu}_w^*)/\hat{\mu}_w^* \), for all windows \( w = 1, \ldots, W \). To further improve the estimate of the dispersion parameter, \( \delta_g \), a smoothing spline technique is used to fit a functional mean-dispersion relationship, \( \delta_g = q(\mu_g) \), by solving the following optimization problem:

\[
\hat{q}_\lambda = \min_{q \in Q} \sum_{w=1}^{W}\left(\delta_w^* - q(\hat{\mu}_w^*)\right)^2 + \lambda \int [q''(t)]^2 \, dt \quad \text{(Eq. 2)}
\]

where \( Q \) is the second-order Sobolev space on \([0, 1]\) containing \( q \), and \( \lambda \) is a smoothing parameter (selected via generalized cross validation). After the fitted curve \( \hat{q}_\lambda \) is obtained as in Eq. 2, the refined estimate of \( \delta_g \) is computed as \( \hat{\delta}_g = \hat{q}_\lambda(\hat{\mu}_g^*) \), \( \forall g \in G_w, w = 1, \ldots, W \). Figure 2 illustrates the functional mean-dispersion relationship and calculation.

![Figure 2](image-url)  
**Figure 2.** Localized, window-specific dispersion estimates as a function of log (mean expression). iDEG partitions the transcriptome into \( W \) equal-sized genomic windows of similar expression size and then calculates the over dispersion parameter relative to the gene’s mean expression. The number of windows \( W \) is a parameter in the iDEG function-call and should be empirically calculated relative to the transcriptome size. After conducting a few numerical studies, we recommend setting \( W=100 \) in order to allow for over dispersion estimates of highly and lowly expressed genes to be representative of their groups.

Summarizing, we get:

a) Initial Estimate of dispersion  
\[
\text{Var}(Y_{gd}) = \sigma^2_{gd} = \mu_{gd} + \delta_{gd} \mu_{gd}^2 \Rightarrow \hat{\delta}_w^* = (\sigma^2_w - \hat{\mu}_w^*)/\hat{\mu}_w^*
\]

b) Refined estimate of dispersion  
\[
\hat{q}_\lambda = \min_{q \in Q} \sum_{w=1}^{W}\left(\delta_w^* - q(\hat{\mu}_w^*)\right)^2 + \lambda \int [q''(t)]^2 \, dt \Rightarrow \delta_g = \hat{q}_\lambda(\hat{\mu}_g^*)
\]
These equations come from the fact that in the negative binomial, variance (equation a) is a function of both mean and dispersion. So, the above equation (a) is rewritten version of the typical variance equation, with the stars and hat superscripts denote that it is now an estimate of the theoretical values for each partitioned window, $w$. As shown in (b), once window-level parameters are estimated (e.g., $\hat{\mu}_w$), then a window-level dispersion parameter is estimated for all genes in that window (e.g., $\hat{\delta}_g = \hat{q}_g(\hat{\mu}_w) \forall g \in G_w$), by fitting refined, functional estimate of dispersion.

**Apply the Variance Stabilizing Transformation ($h(Y_{gd})$) to each gene (Figure 1, Panel 4)**

After fitting $\hat{\delta}_g$, we apply the variance stabilizing transformation $h$ to the counts, $Y_{gd} \forall g \in G_w$:

$$h(Y_{gd}) = \frac{1}{\hat{\delta}_g} \sinh^{-1} \sqrt{Y_{gd} \hat{\delta}_g} + \left(\frac{1}{\sqrt{\hat{\delta}_g}} - 1\right) \sinh^{-1} \sqrt{\frac{Y_{gd} + \hat{\delta}_g}{\hat{\delta}_g}} \forall g \in G_w \text{(Eq. 3)}$$

This transformation\(^{22}\) of the expression $Y_{gd}$ in each window $w$, results in an approximately constant variance across all windows of the transcriptome (Figure 3), regardless of the expression mean, $\mu_{gd}$. That is $h(Y_{gd}) \sim N(h(\mu_{gd}), 1)$, where $d$ specifies if the count comes from sample 1 or 2. Therefore, the difference of the two independent normal random variables (e.g., $h(Y_{g1}) - h(Y_{g2})$ ) approximately follows a common normal distribution with mean 0 and a constant variance: $Z_g = (1/\sqrt{2})\{h(Y_{g1}) - h(Y_{g2})\} \sim N(0,1)$. We suggest replacing $\sqrt{2}$ by a robust estimate of standard deviation (e.g., median absolute deviation)\(^{23}\). In most single-subject analyses, the estimated dispersion parameter, $\hat{\delta}_g$, is small, but when $\hat{\delta}_g \geq 2/3$, the VST $h(Y_{gd})$ is not numerically stable. To avoid this issue, we suggest replacing $h$ with $h^{*} = \left(1/\sqrt{\hat{\delta}_g}\right) \sinh^{-1} \sqrt{Y_{gd} \hat{\delta}_g}$, $g = 1, \cdots, G$; $d = 1,2$. If a negative value of $\hat{\delta}_0$ is obtained, we conservatively set it to zero to assume a larger variance.

**Figure 3. Variance Stabilizing Transformation (VST).** Panel A depicts the raw difference $D_g = Y_{g1} - Y_{g2}$ for 20,000 simulated genes (Methods Simulations), suggesting that the variance of $D_g$ increases as the mean $\mu_{g1}$ increases; hence, there is no uniform cutoff to differentiate DEGs and null genes. Panel B illustrates that, for null genes, VST makes the variance of $D^{*}_g = h(Y_{g1}) - h(Y_{g2})$ constant regardless of mean $\mu_{g1}$.

**Compute the summary statistic for each gene (Figure 1, Panel 5)**

In the context of noisy data and large-scale inference, performing individual tests neglects the parallel structure of RNA-Seq data. Moreover, actual data mean and variance may not be close to their theoretical values of 0 and 1 due to various reasons (e.g., correlation across genes, correlation between samples, or failed mathematical assumptions). Therefore, we estimate an empirical null distribution $N(\mu_0, \sigma_0)$ to test these individual hypotheses.

Since differentially and non-differentially expressed genes generally follow different distributions, the probability density function of $Z_g$, $f(z)$, is naturally modeled by a two-group mixture: $f(z) = \pi_0f_0(z) + \pi_1f_1(z)$. Here, $f_0$ and $f_1$ are the probability density functions of genes in $G'$ and in $G$, while $\pi_0$ and $\pi_1 = 1 - \pi_0$ are their respective membership proportions. We assume a normal distribution following previous work from Dean and Raftery- that applied a two-group mixture model to identify differentially expressed genes, assuming a normal distribution for the null genes and a uniform distribution for the DEGs. However, we relax their assumptions for the marginal distribution and assume an exponential family. We approximate $f(z)$ using a smooth K-parameter exponential

586
family distribution, $f(z) = \exp \left( \sum_{k=0}^{K} \beta_k z^k \right)$, and estimate the parameters ($\beta_0, \beta_1, \cdots \beta_k$) using Efron’s approach.

**Estimate the local false discovery (locfdr) for each and identify DEGs (Figure 1, Panel 6)**

Finally, to control the false discovery rate (fdr), we adopt Efron’s idea to estimate the local fdr (locfdr) using the R package locfdr and estimate $\pi_g$, $f_0(z)$ by maximum likelihood. Efron et al. have shown locfdr’s close connection to the BH false discovery rate procedure; therefore, after estimating locfdr($z_g$), it identifies differentially expressed genes by comparing locfdr($z_g$) to a pre-specified $\alpha$-cutoff value. The final set of differentially expressed genes identified by applying the iDEG procedure is denoted by $\hat{G}$.

**Simulations (Figure 4)**

To compare the performance of iDEG to existing methods – including edgeR, DEGSeq, DESeq, and NOISeq – extensive numerical studies were conducted assuming that RNA-Seq data follow the NB distribution with a varying dispersion parameter $\delta_g$. Of note, these methods assume the NB distribution for data, which is used in the simulation; except for NOISeq that is nonparametric and DEGseq which assumes a binomial distribution. Baseline (normal tissue) and case (tumor sample) transcriptomes are both simulated and assumed to contain $G = 20,000$ genes; the library size of one transcriptome is 1.5 times larger than the other one. The single-subject RNA-Seq datasets are simulated with different percentages of DEGs, including DEG percentage = 5%, 10%, 15%, 20%, 25%, 30%, 35%, and 40%, and also with different window sizes $W = 10, 100, 1000$ (data not shown). For windows size of 10, 100, and 1000 the accuracies of the simulations remained consistent across each of these simulation conditions. $Y_{g1} \sim NB(\mu_{g1}, \delta_g)$ and $Y_{g2} \sim NB(\mu_{g2}, \delta_g)$, where $\mu_{g1}$ follow a discrete uniform over the range $B = \{5, 6, \ldots, 10,000\}$, and the dispersion parameter $\delta_g$ has been set to $\delta_g = 0.005 + 9/(\mu_{g1} + 100)$, per Anders and Huber-. Probabilities for gene expression means, $\mu_{g1}$, are sampled from:

$$P(\mu_{g1}) = (1/500)e^{-(1/500)\times \mu_{g1}}, \quad g = 1, \ldots, 20000.$$  

For the case transcriptome, we set $\mu_{g2} = \mu_{g1}$ for $g \in G'$ and $\mu_{g2}$ follows a Poisson distribution. $d$ is a random variable, and $d = (\mu_{g1} + k \sqrt{\mu_{g1}}) / \mu_{g1}$, $k \sim$ Normal (4,1). Here, $s$ indicates increasing expression ($s = 1$) or decreasing expression ($s = -1$) of a gene in the case transcriptome relative to baseline. Finally, for each gene $g$, we simulate one observation for $Y_{g1}$ and $Y_{g2}$, respectively and test the hypothesis $\mu_{g1} = \mu_{g2}$. At each iteration, a baseline and a case transcriptome are generated to simulate a distinct RNA-Seq dataset. Methods are assessed by their Precision, Recall, and FPR, and $F_1$ score, $F_1 = (2 \cdot \text{precision} \ast \text{recall}) / (\text{precision} + \text{recall})$. The average number of identified DEGs is also reported. Of note, we excluded from the comparison GFOLD, a standard approach that can be applied to TCWR studies, as it only ranks genes without providing a measure of significance, thus prohibiting the accurate comparison with the remaining techniques using precision-recall curves or ROC curves.

**Results**

Figure 4 depicts the accuracies obtained in the simulations while Table 1 contextualizes each method’s performance relative to the number of DEG calls and the number (and %) of genes seeded as DEGs. Of note, we have also conducted complementary analyses with a Poisson distribution and showed similar ordering of accuracies between the evaluated methods (data not shown). For windows size of 10, 100, and 1000 the accuracies of the simulations remained consistent (data not shown) and opted for setting $W = 100$ to balance computation time and parameter estimation robustness. As seen in Table 1, NOISeq-sim, edgeR, DEGseq attain a high precision (defined as > 90% precision) across all simulation conditions (5% through 40% DEGs seeded) at the expense of lower recall and a large number of false positives. For example, as seen in Table 1, 5% DEGs, NOISeq-sim, edgeR, and DEGseq all result in a larger of false positives than there are actual seeded genes. Conversely, iDEG attains a high precision (defined as >90%) at the expense of making a smaller number of DEG calls, thus attaining lower recall. The $F_1$ score shown in Fig 4-C is the harmonic mean of the Precision and Recall metrics, aggregating the precision-recall trade-offs made by individual techniques into a single technique. Although iDEG never attains as high a recall as DEGseq, edgeR, and NOISeq-sim, it better balances its precision-recall trade-off into a higher overall $F_1$ score at FDRs<20%, while NOISeq does better at FDR>30%, and the two methods show similar $F_1$ scores at 20%<FDR<30%. Of note, DESeq failed to make any DEG calls across the majority of the simulation conditions (since it either produced “0” or “1” fdr-adjusted probability predictions), preventing us from evaluating their performance at any reasonable false discovery cutoff.
Figure 4. Performance results. NOISeq-sim’s and iDEG’s F-scores are more accurate than that of other methods at 5%<DEGs<20% (iDEG) and 30%<DEGs<40% (NOISeq-sim). At DEG=5% and FDR<10%, iDEG provides an interesting compromise between precision and recall, while NOISeq provides a better compromise when the percentage of DEGs is higher than 30%. Panels A & B: Precision recall curve at 10% FDR for 1,000 and 8,000 seeded DEGs among 20,000 transcripts, respectively. Panel C: F₁ scores. Average F₁ scores resulting from 1,000 repeated experiments with vertical bars representing one standard deviation.

Discussion

No single method has emerged as the optimal approach for all conditions. Low expression levels are extremely susceptible to unstable fold-change estimation, as a 5-fold increase from 2 to 10 counts on a dynamic range of 0 to 100,000 should not be treated equivalently to that between 10,000 and 50,000. Standard practice filters out genes with counts below a certain threshold (typically 5 or 10). However, this solution does not address fold change (FC) inflation above the threshold (e.g., FC>2 at 15 counts), nor how to compare distinct FCs at different expression levels. Alternatively, favoring absolute count difference to identify DEGs leads to a systemic bias towards genes with high expression. Conversely, favoring FC results in a systemic bias towards lowly expressed genes. Either of these solutions yields higher false positive rates. For DEGs<30%, the variance stabilization within partitioned windows proposed in iDEG is shown to address this dilemma of dealing with fold change inflation by comparing FC values relative to their expression levels, perhaps because conventional approaches impose stringent data assumptions that may compromise downstream inferential processes.
Table 1. Performance results of TCWR simulations. At different percentage of DEGs in TCWR simulations, distinct methods obtain the best precision and recall, with iDEG, NOISeq and edgeR producing the best combinations of precision and recall. Of note, edgeR, DEGseq, and DESeq were not designed nor validated for studies without replicates; however, their authors proposed to utilize them in these conditions by defining specific parameters. NOISeq-sim offers high recall and precision with DEGs=40% i.e. when 8,000 genes are dysregulated among 20,000. On the other hand, iDEG obtains high precision with moderate to high recall in all conditions. EdgeR provides moderate precision with very high recall for DEGs>20%.

<table>
<thead>
<tr>
<th>Proportions of DEGs seeded</th>
<th>Method</th>
<th>Precision</th>
<th>Recall (TPR)</th>
<th>FP</th>
<th>Predicted DEGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000 out of 20,000 (5% of genes)</td>
<td>iDEG</td>
<td>0.93</td>
<td>0.65</td>
<td>57</td>
<td>700</td>
</tr>
<tr>
<td></td>
<td>edgeR</td>
<td>0.31</td>
<td>0.96</td>
<td>2,090</td>
<td>3,119</td>
</tr>
<tr>
<td></td>
<td>NOISeq-sim</td>
<td>0.412</td>
<td>0.89</td>
<td>1140</td>
<td>2,163</td>
</tr>
<tr>
<td></td>
<td>DESeq</td>
<td>1.0</td>
<td>0.16</td>
<td>0</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>DEGseq</td>
<td>0.086</td>
<td>0.99</td>
<td>10,450</td>
<td>11,397</td>
</tr>
<tr>
<td>4,000 out of 20,000 (20% of genes)</td>
<td>iDEG</td>
<td>0.97</td>
<td>0.76</td>
<td>112</td>
<td>3,136</td>
</tr>
<tr>
<td></td>
<td>edgeR</td>
<td>0.60</td>
<td>0.97</td>
<td>2,560</td>
<td>6,405</td>
</tr>
<tr>
<td></td>
<td>NOISeq-sim</td>
<td>0.747</td>
<td>0.91</td>
<td>1120</td>
<td>4,897</td>
</tr>
<tr>
<td></td>
<td>DESeq</td>
<td>Not applicable</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DEGseq</td>
<td>0.3</td>
<td>0.99</td>
<td>9,280</td>
<td>13,160</td>
</tr>
<tr>
<td>8,000 out of 20,000 (40% of genes)</td>
<td>iDEG</td>
<td>0.98</td>
<td>0.80</td>
<td>120</td>
<td>6,494</td>
</tr>
<tr>
<td></td>
<td>edgeR</td>
<td>0.77</td>
<td>0.97</td>
<td>2,400</td>
<td>10,145</td>
</tr>
<tr>
<td></td>
<td>NOISeq-sim</td>
<td>0.893</td>
<td>0.91</td>
<td>840</td>
<td>8,177</td>
</tr>
<tr>
<td></td>
<td>DESeq</td>
<td>Not applicable</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DEGseq</td>
<td>0.51</td>
<td>0.99</td>
<td>7,560</td>
<td>15,455</td>
</tr>
</tbody>
</table>

As we proceeded to validating iDEG in biologic or clinical datasets, a review of literature identified few candidate datasets that comprised targeted mutations over an isogenic background and yield high DEG rates (e.g., DEGs>50%) that did not reflect rates expected in clinical care. In addition, the state of the art in generating reference standard consisted in comparing one method against itself as the overlap of DEGs across conventional methods was low in spite of 30 replicates in isogenic conditions. Because of these two considerations, we decided to publish the results of a comprehensive improvement in reference standard generation as a companion paper33. We have thus generated multiple distinct reference standards (one per conventional method) and developed a “fair” evaluation of methods to identify DEGs in paired conditions without replicates using biological datasets (each method is compared to all other methods but not itself)33. This companion biological paper33 is limited to datasets with high DEGs as no reference datasets were available for low DEGs conditions, while the current simulation explores both low and moderate DEGs levels. We will extend window-level approximations for all techniques considered in our companion paper33 in future studies.

We note several limitations to the current study. First, conventional techniques were not explicitly designed for absence of replicates and are tested in those conditions. In addition, each method assumes some distribution (DEGseq assumes a binomial distribution; iDEG, edgeR, and DESeq assume a Negative Binomial distribution; and NOISeq is non-parametric). Since the distributional form in real-data is never truly known (only approximated), simulating a transcriptome necessarily entails distributional assumptions in every simulation study, which limits its generalizability to real studies and inherently may favor some methods over the others. In cases where biological replicates exist, the existing conventional methods are better powered and more appropriate for inference as their assumed distributions and parameters can be estimated. In cases where biological replicates are unfeasible, the assumptions of these conventional methods are violated, and their performance is not guaranteed. In some cases, some DEG detection techniques actually cannot produce a DEG call in TCWR. This limits the number of techniques available for comparison in our simulation study. Moreover, as seen with DESeq, DEG techniques designed for replicated studies are not necessarily fully operational or effective in TCWR, therefore it is not necessarily recommended to pick an arbitrary DEG technique and use it in non-replicated TCWR studies. We conducted the simulations against these methods to illustrate the need for new approaches to study single-subject transcripts in TCWR conditions. In addition, a true gold standard to evaluate iDEG and other methods is not as simple as obtaining replicates and running conventional methods as pointed out by recent papers34,35.
Conclusion

Over the past decade, state-of-the-art techniques in RNA-Seq data analysis have delivered powerful new tools for extending large-scale inference to small-sample settings. The primary goal of iDEG is not to replace these, but rather to expand the scope of RNA-Seq studies into the single-subject, single-replicate realm and provide novel research opportunities and test methods for controlling fold change inflation at low expression ranges. In iDEG, we have shown the novelty of window partitioning to borrow localized distribution information across genes, and its improved accuracy over alternate methods in low DEG conditions (DEG<20%). Furthermore, this approach could potentially be applied to improve the accuracy of existing parametric and non-parametric differential expression tools. In future studies, we envision to i) extend the window partitioning component of iDEG into other techniques, ii) to locally identify differentially expressed pathways (by incorporating ontologies and knowledge graphs), and iv) to apply it to other ‘omics measures, (e.g., metabolomics, proteomics, etc.).

List of Abbreviations

- DEGs = differentially expressed genes
- FC = fold change
- FCI = fold change inflation
- FDR = false discovery rate
- iDEG = individualized Differentially Expressed Genes
- locfdr: local false discovery rate
- NB = Negative Binomial distribution
- RNA-Seq = RNA Sequencing
- TCWR = two conditions without replicates
- VST = Variance Stabilizing Transformation

Declarations

Availability of data and material

Software is available at http://www.lussiergroup.org/publications/iDEG

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, NIH (U01AI122275, CA023074, 1UG3OD023171, IS10RR029030). This article did not receive private sponsorship.

Authors' contributions

Conceived the study: QL, HHZ, YAL. Experimental design and analysis: SRZ, QL, DA, YAL, HHZ. Manuscript writing: SRZ, QL, YAL, JB Figures: SRZ, CK, YAL. Interpretation: SRZ, QL, DA, JB, FV, HL, YAL. All of the authors have read and approved the final manuscript.

References


High Performance Computing on Flat FHIR Files
Created with the New SMART/HL7 Bulk Data Access Standard

Dianbo Liu, PhD,1,2 Ricky Sahu,3 Vlad Ignatov,1 Dan Gottlieb, MPA,4 Kenneth D. Mandl, MD, MPH1,4
1Computational Health informatics Program, Boston Children's Hospital, Boston, MA, USA; 2 Department of Pediatrics, Harvard Medical School, Boston, MA, USA, 3 1upHealth, Boston, MA, USA, 4Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.

This project was supported as a Leading Edge Acceleration Project (LEAP) under contract 90AX0022/01-00 from the Office of the National Coordinator of Health Information Technology

Abstract
The FHIR Bulk Data API is designed to create a uniform capability for population-level exports from clinical systems, into a file format often referred to as “Flat-FHIR.” Leveraging the SMART backend services authentication and authorization profile, the approach enables healthcare providers and organizations to define and access cohorts from electronic health records and payor claims data with “push button” simplicity—a substantial advance over the current state, where each site of care needs highly skilled extraction transform and load (ETL) efforts.

Introduction
Healthcare providers and organizations using electronic health records do not currently have access to turnkey, standardized methods to export bulk data on large cohorts of individuals. Instead, local teams expert in extraction, transformation and loading of data and in medical terminologies are needed. Yet, there are myriad use cases for population datasets; a healthcare organization may want to frequently upload patients’ data from an electronic medical record system to a research data warehouse. To date, the most widely used FHIR API, SMART on FHIR, is targeted at retrieving data on a single patient at a time. Using this API for population data exports, requires thousands or even millions of queries. In order to facilitate population-level analysis, the SMART project is collaborating with Health Level Seven International (HL7), to extend the FHIR API with a Bulk Data Access method that can be used for large scale sharing from any healthcare system with FHIR implemented as interoperability layer.

The FHIR Bulk Data Access API used a NDJSON-based data format as the default export format for the FHIR data. The resulting file is referred to “Flat FHIR.” JavaScript Object Notation (JSON) is a text data format widely used and with long history. Newline delimited JSON (NDJSON) format makes use of new line to separate different values where each line is a JSON object removing the necessity of loading the entire data set into memory when reading or writing values. When used for FHIR data, each line represents a separate record such as an immunization record of a child or demographic information of a patient.

As the standard becomes more widely adopted and the file size of exports grow, analysts will need to select a data format for efficient computation. To plan for robust analytics on Flat FHIR exports, we seek to compare computation directly on the native NDJSON format with two others—AVRO and Parquet. AVRO is a binary data format, which stores both the data definition and data in the same field and was originally developed for Hadoop. A key feature of AVRO is the support of evolutionary data schemas. Parquet is a binary column oriented data storage format, which means the data were stored by columns instead of rows by the management system. Implemented using record-shredding and assembly algorithms, Parquet files can be efficiently compressed efficiently.

Methods
Dataset
Synthetic data for 1,400,000 immunizations among 152,072 patients were generated using MITRE’s Synthea. The data includes patient ID, diagnosis, medical encounters, social determinants of health, medications and immunization records.
Computational environment
Analyses were conducted using Spark 2.4.0 in Scala 2.11.12 with the spark-avro_2.11 module. The processor was a 2.8 GHz Intel Core i7 cpu with 16 GB 1600 MHz DDR3 RAM, and a solid state hard drive. Java version 1.8.0_202 was used when needed for Scala, Python or Spark functions.

Format conversion and file reading and writing
Flat FHIR file Data in an NDJSON format were loaded into Spark and then converted to AVRO (using Spark-avro_2.11 module) or Parquet (read in to Spark as Spark as dataframes and then written into files) NDJSON, AVRO and Parquet were all read into SQL queries. A series of different SQL queries were tested on all three data formats in the Spark environment. The execution time calculated for each query (Table 1) includes both the query and visualization of head of the query results.

Machine learning platform compatibility
All data loaded into Python can be converted to numpy array or dataframe for use in the machine learning platforms Tensorflow, PyTorch and PySpark. We defined compatibility as the ability to directly load and stream untransformed data into the into machine learning algorithms for training or inference purposes. Petastorm is a Python library developed for machine learning models training directly from Parquet file format data. Petastorm supports Tensorflow, PyTorch and PySpark. When looking at compatibility with existing cloud based machine learning engine, we considered the data format if it can be directly taken as input file with conversion based on the user guide or the description on the cloud service website, and without conversion to comma separated value (CSV) or other formats.

Results
Data File
The Flat FHIR file size was 771 MB in NDJSON format. An example of an individual patient record is shown in Figure 1.

Figure 1. An example of the immunization record of a synthetic patient.

Storage size and write/read speed
Reading and writing speeds and storage requirements were compared for native NDJSON, and AVRO and Parquet formats. On average, one patient immunization record occupies 10.3 bytes in an NDJSON file, 0.7 bytes in AVRO and 1.2 bytes in Parquet (Figure 2A). The average reading speed per immunization record is 1.2 µs for NDJSON, 0.2 µs for AVRO, and 0.3µ seconds for Parquet. On average it took 4.6 µs, 5.2 µs, and 5.1 µs to write a immunization record into NDJSON, AVRO and Parquet file respectively.
Figure 2. Storage size, reading and writing speed of in NDJSON, Flat-FHIR AVRO and Parquet formats. (A) Storage sizes (bytes/immunization record). (B) Reading/writing speed (µs/immunization record) SQL query speed

The first query extracted all the information from the data file. The second query extracted all plain descriptions of immunization such as “Influenza, seasonal, injectable, preservative free” from a data file. The third query retrieved patient IDs for patients who did not receive the vaccine during the visit. The fourth query joined the immunization and patient demographic information. On average, SQL queries on data in AVRO and Parquet format execute faster than in NDJSON format.

<table>
<thead>
<tr>
<th>Queries (in seconds)</th>
<th>NDJSON</th>
<th>AVRO</th>
<th>Parquet</th>
</tr>
</thead>
<tbody>
<tr>
<td>select * from table_name</td>
<td>10.3</td>
<td>0.7</td>
<td>1.2</td>
</tr>
<tr>
<td>select vaccineCode.coding.display from table_name</td>
<td>3.8</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>select patient FROM table_name where wasNotGiven=true</td>
<td>5.6</td>
<td>2.6</td>
<td>2.0</td>
</tr>
<tr>
<td>left join immunization table and patient information table</td>
<td>8.2</td>
<td>5.6</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Machine learning platform compatibility
Compatibility of NDJSON, AVRO and Parquet with the Tensorflow, Pytorch and PySpark were examined. NDJSON is compatible with PySpark and Pytorch but not Tensorflow. AVRO works with only PySpark. Parquet is compatible with all three machine learning platform using Petastorm library (Table 2).

In addition to machine learning platforms that run locally, we tested the, compatibility of data formats with four widely used cloud based machine learning engines, Amazon cloud service machine learning engine, Google Cloud Machine Learning, Microsoft Azure, and IBM cloud machine learning. While Google Cloud machine learning engine works with all three data formats, all the other three engines do not work with NDJSON, AVRO or Parquet (Table 2).
Table 2. Compatibility of Flat FHIR file formats with machine learning platforms and cloud based machine learning services

<table>
<thead>
<tr>
<th></th>
<th>NDJSON</th>
<th>AVRO</th>
<th>Parquet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compatibility with ML platforms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tensorflow</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pytorch</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PySpark</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Compatibility with Cloud based ML engines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon machine learning</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Google Cloud AI</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Azure machine learning</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>IBM machine learning</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Discussion and Conclusion

The FHIR and SMART-on-FHIR standards improve interoperability between different data silos and make cross-silos app development possible. The FHIR Bulk Data API may similarly improve population health analysis. To meet different requirements of clinical practices and medical research, different data formats should be considered based on a project’s requirements.

NDJSON, as the default data format used by the FHIR Bulk Data API, has a simple structure, higher writing efficiency and is human readable. However, storage size and query speed are sacrificed under some analytic conditions. AVRO and Parquet have higher storage efficiency and faster query speed, but are slower to writing and are not human readable. Parquet is most compatible with the machine learning platforms we evaluated. Given these findings, our tentative recommendation is to use the NDJSON NDJSON format for data exchange, and performing analytic tasks on small data sets, but transform the data into Parquet format for large scale analytic projects. Given the costs of using cloud hosted environments, these efficiencies could lead to millions of dollars of savings.

References


Learning Hierarchical Representations of Electronic Health Records for Clinical Outcome Prediction

Luchen Liu¹, Haoran Li¹, Zhiting Hu², Haoran Shi¹, Zichang Wang¹, Jian Tang, PhD³,⁴, Ming Zhang*, PhD¹
¹Department of Computer Science, Peking University, Beijing, China
²Carnegie Mellon University, Pittsburgh, PA, US
³Mila - Québec AI Institute, Montréal, Québec, Canada
⁴HEC Montréal, Montréal, Québec, Canada

Abstract

Clinical outcome prediction based on Electronic Health Record (EHR) helps enable early interventions for high-risk patients, and is thus a central task for smart healthcare. Conventional deep sequential models fail to capture the rich temporal patterns encoded in the long and irregular clinical event sequences in EHR. We make the observation that clinical events at a long time scale exhibit strong temporal patterns, while events within a short time period tend to be disordered co-occurrence. We thus propose differentiated mechanisms to model clinical events at different time scales. Our model learns hierarchical representations of event sequences, to adaptively distinguish between short-range and long-range events, and accurately capture their core temporal dependencies. Experimental results on real clinical data show that our model greatly improves over previous state-of-the-art models, achieving AUC scores of 0.94 and 0.90 for predicting death and ICU admission, respectively. Our model also successfully identifies important events for different clinical outcome prediction tasks.

1 Introduction

The ever-growing massive Electronic Health Records (EHR) data expose an opportunity for large-scale data-driven health analysis and intelligent medical care. Predicting clinical outcomes, such as death and intensive care unit (ICU) admission, plays an important role in improving the performance of healthcare systems. For instance, accurate clinical prediction based on patients existing medical records can enable advanced and timely medical intervention.

Clinical outcome prediction is challenging because it is hard to utilize rich temporal information encoded in the sequence of clinical events in EHR data.¹ In particular, EHR data usually consist of clinical events with irregular intervals² from heterogeneous sources, including patient health features (vital sign measurements, laboratory test results, etc), medical interventions (procedures, drug inputs, etc), and expert judgments (diagnoses, notes, etc).³ The temporal order of these events is critical for predicting clinical outcome. For example, patient health features can be affected by previous medical interventions, and in turn determine subsequent medical interventions through expert judgments.

Conventional approaches that directly apply classic deep sequential models, such as recurrent neural networks⁴,⁵ (RNN) and convolutional neural networks⁶ (CNN), usually fail to capture temporal dependencies in such long irregular event sequences,⁷,⁸ as long-term dependencies can easily exceed the modeling capacity. To handle irregularly timed events, some extensions of the classic models have been developed, such as time-aware RNN⁹ or CNN.¹⁰ However, their performance is still largely unsatisfying due to the limited ability to capture long-term dependencies.

This work aims to address the above challenges. We first make a key observation that, though the clinical events in EHR data can exhibit strong temporal patterns at a long time scale, the events occurring within a short time period usually do not have a definite order. Specifically, unlike word sequences in natural languages where word tokens are ordered by grammar rules, clinical events recorded in a short period of time are instead a series of events, such as clinical laboratory test results, that only reflect the patients status in different views. Therefore, direct temporal modeling of such short-range events as in previous work can introduce noise and harm the temporal predictive performance. Instead, local dependencies of these events should be modeled as event co-occurrence, and we can further select critical events from each of these short-range event groups as the basis for modeling the real temporal dependencies at

*Corresponding author.
a long time scale. A key difficulty to this end is that the criterion for distinguishing long-term temporal dependencies from the local co-occurrence of critical events in a short range can vary across different diseases and phases, especially in the irregular EHR data.11,12

To address the difficulties mentioned above, we propose a hierarchical neural network for clinical outcome prediction. Specifically, we adaptively segment irregular event sequences into sequential groups to distinguish short-range co-occurrence and long-term temporal dependencies as well as to learn hierarchical representations of the event sequence. At the low level, the model automatically identifies critical clinical events in each group and aggregates the events to form event group representations. At the high level, meaningful long-term dependencies of clinical event groups are captured in a sequence representation vector by a recurrent neural network. Compared to traditional methods, the proposed method has several advantages:

- Our model can deal with the temporal irregularity of clinical event sequences by adaptively segmenting an event sequence into sequential groups.
- Our model learns a hierarchical representation of long and irregular event sequences to capture long-term dependencies of clinical events.
- The model is capable of discovering critical event groups as well as critical events in each group, through a temporal attention and event attention mechanism. This provides useful clinical insights for accurate prediction.

2 Related Works

2.1 Modeling EHR Data

Most existing works based on EHR data have either focused on stationary clinical text and images, or ignored irregular time intervals of temporal clinical events. For example, previous work trained the semantic embeddings for the categories of clinical events for adverse drug event detection, or proposed a multi-view learning method that generalizes Canonical Correlation Analysis for an arbitrary collection of matrices involving missing data. These works make predictions based on the clinical events with regular time intervals, and cannot distinguish short-range order from long-term temporal order of different diseases and patients. Our work addresses the issue by adaptive segmentation of clinical event sequences.

As long-term temporal dependencies are hard to capture, many works use a small subset of the whole EHR information, to avoid dealing with the long clinical event sequences. Some works select a subset of the numerical clinical features (the numerical attributes of clinical events) in the EHR data according to the expertise of clinicians. For instance, Yoon only uses a set of 21 (temporal) physiological streams comprising a set of 11 vital signs and 10 lab test scores to predict ICU admission. Some works used graphical models to model patients health status. Some techniques transform selected 99 time series features of all the EHR data into a new latent space using the hyper-parameters of multi-task GP (MTGP) models to model patient similarity. Recently, RETAIN used two reversed recursive neural networks (RNN) generating attention variables of sequential international disease classification (ICD) code groups for the prediction. However, the codes are grouped by the fixed-length time slots for distinct patients and diseases, and local dependencies and long-term dependencies may be mixed up. But these works can lose significant information, due to the expert bias when selecting a limit fraction of all clinical features in EHR as the input of the models, and fail to provide new data-driven insights for better healthcare.

2.2 Clinical Outcome Prediction

The clinical outcome prediction problem is studied by many works. However many of them cannot take advantage of the temporal information in EHR data for prediction. Some of these studies used latent variable models to decompose bag-of-words free-text extracted from clinical event descriptions into meaningful features to predict patient mortality. “Deep patient” arranged all clinical descriptors (features) in a period of time in a sparse vector without temporal information and trained the deep representation of patients with a 3 layer denoising autoencoder for diagnosis. Some work studied how to diagnose and predict Alzheimers disease (AD) with a hybrid manifold learning for non-temporal clinical feature embedding and the bootstrap aggregating (Bagging) algorithm. There is also a work model EHR
data by factorizing the medical feature matrix into a latent medical concept mapping matrix and a concept value evolution matrix, and then they averaged all vectors in the evolution matrix to predict heart failure.\textsuperscript{27} Our model learns the hierarchical representations of clinical event sequences to utilize the temporal information for clinical outcome prediction.

3 Data and Task Descriptions

We give the notations and data descriptions of the predictive tasks in the following.

Clinical Events in EHR

A clinical event is a record in the database of EHR, which describes a clinical activity of a particular patient at a certain time. The events can be measurements of vital signals, injection of drugs, results of laboratory tests, and so on, which are summarized in Table 1. Each clinical event has some attributes, including categorical attributes and numerical attributes. For example, the lab test event $e_t$ has 2 categorical features and 1 numerical feature: $e_t = \{\text{LabItem} : \text{Cholesterol}, \text{AbnormalLabel} : \text{Abnormal}, \text{ResultIndex} : 51\mu/L\}$. The meaning of this event is the result of the Cholesterol test is $51\mu/L$, which reflects an abnormal health status. An episode of a patient EHR data is a clinical event sequence, which may consist of hundreds of clinical events.

Clinical Outcome Prediction

Clinical outcome prediction is to dynamically predict whether a clinical outcome will happen in 24 hours based on an episode of a patient. We aim to dynamically predict two outcomes in this work. In the first “death prediction task”, the outcome is death in hospital or discharge to home. In the second “ICU admission prediction task”, the outcome is clinical deterioration (need to be immediately transformed to ICU), or stable clinical status.

Patient Cohort Setup

We set up two datasets from one real clinical data source, MIMIC-III\textsuperscript{3} (Medical Information Mart for Intensive Care III), which is a large, freely-available database comprising de-identified health-related data associated with over forty thousand patients who stayed in intensive care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

We extract 18192 kinds of clinical events with their attributes from the database to get event sequences of patients (the events with top frequency are listed in Table 1). The events whose frequency is less than 2500 are dropped out. And we also drop out the admissions, of which the time span from the beginning to the target clinical outcome is less than 36 hours. Each input of a sample is an episode of a patient clinical event sequence 24 hours before the target outcome. The statistics of final clinical event sequences in the two tasks are summarized in Table 2.

4 Methodology

In this section, we introduce the technical details of our proposed model. Our model first segments the whole clinical event sequence into several event groups via the adaptive event sequence segmentation module. Then the model learns hierarchical representations of event sequences with both event attention and temporal attention mechanisms. The architecture of our model is illustrated in Figure 1.

4.1 Adaptive Segmentation

To distinguish long-term temporal dependencies from co-occurrence of important events in short range, we adaptively segment an event sequence for a patient into sequential groups, according to the irregular record time of events. As events in the same group are exchangeable, sequential groups can avoid the influence of the short-range order noisy in clinical events. Moreover, sequential groups reduce the length of the sequences fed to RNN, which makes capturing long-term temporal dependencies easier.
Table 1: Statistics of High Frequency Examples in Different Types of Clinical Events

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Event Name</th>
<th>Frequency</th>
<th>Coverage</th>
<th>Frequency per Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chart events</strong></td>
<td>Heart Rate</td>
<td>5171250 (0.01%)</td>
<td>0.64 (0.16%)</td>
<td>173.4</td>
</tr>
<tr>
<td></td>
<td>SpO2</td>
<td>3410702 (0.01%)</td>
<td>0.479 (0.33%)</td>
<td>153.0</td>
</tr>
<tr>
<td><strong>Input events</strong></td>
<td>0.9% Normal Saline</td>
<td>2363812 (0.05%)</td>
<td>0.393 (0.66%)</td>
<td>129.2</td>
</tr>
<tr>
<td></td>
<td>Propofol</td>
<td>369103 (0.81%)</td>
<td>0.217 (1.45%)</td>
<td>36.5</td>
</tr>
<tr>
<td><strong>Lab events</strong></td>
<td>Hematocrit</td>
<td>881846 (0.22%)</td>
<td>0.976 (0.01%)</td>
<td>19.4</td>
</tr>
<tr>
<td></td>
<td>Potassium</td>
<td>845825 (0.23%)</td>
<td>0.886 (0.05%)</td>
<td>20.5</td>
</tr>
<tr>
<td><strong>Procedure events</strong></td>
<td>Chest X-Ray</td>
<td>32723 (3.2%)</td>
<td>0.204 (1.52%)</td>
<td>3.44</td>
</tr>
<tr>
<td></td>
<td>EKG</td>
<td>13962 (4.35%)</td>
<td>0.167 (1.82%)</td>
<td>1.79</td>
</tr>
<tr>
<td><strong>Output events</strong></td>
<td>Chest Tubes CTICU CT 1</td>
<td>151766 (1.57%)</td>
<td>0.098 (2.67%)</td>
<td>33.2</td>
</tr>
<tr>
<td></td>
<td>Urine</td>
<td>107465 (1.93%)</td>
<td>0.075 (3.05%)</td>
<td>30.8</td>
</tr>
</tbody>
</table>

Table 2: Statistics of the datasets (the percentage in the second column is the positive sample rate)

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of samples</th>
<th># of events</th>
<th>Avg timespan</th>
</tr>
</thead>
<tbody>
<tr>
<td>death</td>
<td>24301 (8%)</td>
<td>20290879</td>
<td>3d 15h 58m</td>
</tr>
<tr>
<td>ICU admission</td>
<td>19451 (21%)</td>
<td>14515198</td>
<td>4d 18h 31m</td>
</tr>
</tbody>
</table>

We find segmentation points of an event sequence by minimizing the maximum time span of the resulting segments. Formally, given a event sequence \( \{e_t\} \), \( k-1 \) segmentation points can split the sequence into \( k \) groups \( \{G_i\}_{i=1}^k \), where the event group \( G_i = \{e_t | t'_i \leq e_t.time \leq t'_i+1\} \) is an episode of clinical events from time \( t'_i \) to time \( t'_i+1 \), where \( e_t.time \) is the record time of the event \( e_t \). And the time span of a group is defined as the time difference of the last event and the first in the group, namely \( \text{span}(G_i) = \max_{e_k,e_j \in G_i} \{e_k.time - e_j.time\} \). So the optimal choice of the segmentation points can be found by minimizing the following:

\[
t'_1, ..., t'_{k+1} = \arg\min_{t'_1, ..., t'_{k+1}, k \leq M} \max_{1 \leq i \leq k} \{\text{span}(G_i)\}
\]

where \( M \) is the max number of groups and \( k \leq M \) is the constraint to avoid the segmentation too fine-grained.

The adaptive segmentation is designed in a way of the combination of greedy method and binary search. We binary search the minimal upper bound of the maximum time span of all groups. And we then verify the searched upper bound of time spans by trying to greedily construct a solution satisfying the constraints of \( M \) and the time span upper bound. The time complexity of the algorithm is \( O(L \times \log T) \), where \( L \) is the length of the event sequence and \( T \) is the difference of end time and start time of the sequence. We can regard this algorithm as an algorithm of linear complexity with a big consistent coefficient.
Figure 1: The overall architecture of our model. The original irregular event sequence is segmented into sequential event groups by the adaptive segmentation module. Then our model learns the hierarchical representation of the sequential event groups. In the low-level representations, each event group is represented as a vector $g_i$ by the event attention. In the high-level representations, the embedded sequential groups are modeled by general recurrent units (GRU) with inter-group temporal attention.

Clinical events have many attributes, which are not considered in previous popular methods, such as word2vec, GloVe and so on. To represent clinical events with their attributes, we embed each clinical event $e_t$ into the low dimension space as a vector $v_t$ in the way described in the previous work. The representing vector $v_t \in \mathbb{R}^N$ (where $N$ is the event embedding dimension) is the sum of event type vector (as basic event information) and event attribute encoding vector (as the description of event feature).

4.2 Hierarchical Representations with Attention Mechanisms

Based on the sequential event groups, the model can learn hierarchical representations to capture long-term temporal dependencies. In the low-level model, the model automatically identifies critical clinical events in each group via event attention mechanism and aggregates the events to form event group representations. In the high-level model, the meaningful long-term temporal dependencies of clinical event groups are captured by a recurrent neural network with temporal attention mechanism in the sequence representation. The hierarchical representations help to learn long-term temporal dependencies in the original event sequences.

4.2.1 Event Group Representation

To select the significant events in each group and compact events in the same group into one vector as the event group representation, the event attention mechanism are added in the low-level model.

Given sequential groups produced by the adaptive segmentation module $\{G_i\}_{i=1}^T$, where $G_i = \{v_{i1}, ..., v_{ini}\}$, attention score of each event in the group is calculated by the event attention mechanism. The scalars $\alpha_{i}^{t}$ are the event attention weights that govern the influence of event embeddings $v_{i1}, ..., v_{ini}$ in the group $G_i$.

We use a multi-layer perceptron (MLP) with one hidden layer to generate $\alpha_{i}^{t}$ based on the event embedding vector $v_{i}^{t}$ and the hidden state of the previous time as follows:
\[ q_i^t = w_q \times \tanh(W_e \times v_i^t + W_h \times h_{i-1} + b_h) \]

where \( 1 \leq i \leq T, 1 \leq t \leq n_i \).

\[ \alpha_1^i, ..., \alpha_{n_i}^i = \text{softmax}(q_1^i, ..., q_{n_i}^i) \text{ for } 1 \leq i \leq T \]

where \( h_{i-1} \in \mathbb{R}^S \) is the hidden state of the previous gated recurrent units (GRU),\(^29\) (which will be described in the following section) \( q_i^t \) is the latent layer of the event \( e_i \) at group \( i \) and \( W_e \in \mathbb{R}^{H \times N}, W_h \in \mathbb{R}^{H \times S}, b_h \) and \( w_q \in \mathbb{R}^H \), are parameters to learn. Notice that \( H \) is the hidden layer dimension and \( S \) is the GRU hidden state dimension.

The resulting attention scores \( \alpha \) reflect the importance of each event in a group according to the temporal context of the group. Events in the \( i \)-th group are weighted averaged with \( \alpha \) to get the group representation \( g_i \in \mathbb{R}^N \) as the input to the \( i \)-th GRU unit.

\[ g_i = \sum_{t=1}^{n_i} \alpha_i^t \times v_i^t \]

4.2.2 Sequence Representation

To spot the critical phases over the sequence for the final decision and capture long-term temporal dependency of event groups, gated recurrent units (GRU)\(^29\) equipped with temporal attention mechanism is employed as the high-level model.

\[ h_i = \text{GRU}(h_{i-1}, g_i, \theta) \]

where the function \( \text{GRU}(\cdot) \) represents the recurrent unit, which use the previous hidden state \( h_{i-1} \) and current input vector \( g_i \) to update the hidden state. And \( \theta \) represents all the parameters of GRU.

The vector \( \beta = (\beta_1, ..., \beta_T) \) contains the temporal attention weights of each group in the sequence. And we use a fully connected feedforward network to generate \( \beta \) from the output of GRU at each time as follow:

\[ \beta = \text{softmax}(w_{\text{temporal}} \times O_T) \]

where \( O_T = (h_1, h_2, ..., h_T) \in \mathbb{R}^{S \times T} \) is the output matrix and \( w_{\text{temporal}} \in \mathbb{R}^S \) is a vector of parameters to learn.

The sequence representation \( s \) is the weighted average of the output matrix \( O_T \). We use \( s \) to predict the true label \( \hat{y}_T \) of the sequences.

\[ s = O_T \times \beta \]

\[ y_T = \text{sigmoid}(w_p \times s + b_p) \]

where \( w_p \in \mathbb{R}^S \) and \( b_p \) are parameters to learn.

The cross-entropy loss function is used to calculate the classification loss of each sample as follows:

\[ \text{Loss}([e_t], \hat{y}_T) = \hat{y}_T \times \ln y_T + (1 - \hat{y}_T) \times \ln(1 - y_T) \]

where \( \{e_t\} \) is the input event sequence and \( \hat{y}_T \) is the label indicating whether the clinical outcome happens. And we can sum up the losses of all the samples in one mini-batch to get the total loss for back propagation.

5 Results and Discussions

5.1 Comparison Methods and Settings

We compare our model with popular models in the literature, which include bag-of-words vector classifiers (i.e. support vector machine\(^30\) (SVM), logistic regression\(^31\) (LR), random forest\(^32\) (RF)) and deep sequential models, such as RETAIN\(^23\) and RNN\(^12\) (implemented with GRU).

Due to the fact that SVM, LR, RF cannot handle a sequence input, the event sequence is compressed into a 0-1 vector in which the \( i \)-th element indicates whether the \( i \)-th event happens, and then fed into SVM, LR, or RF to make the outcome prediction. We implement these bag-of-words vector classifiers using scikit-learn (https://scikit-learn.org).

---

602
Table 3: performance of different models on death and ICU admission prediction tasks

<table>
<thead>
<tr>
<th>Models</th>
<th>Death AUC</th>
<th>Death AUPRC</th>
<th>ICU admission AUC</th>
<th>ICU admission AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.7523</td>
<td>0.5154</td>
<td>0.7973</td>
<td>0.7074</td>
</tr>
<tr>
<td>LR</td>
<td>0.8843</td>
<td>0.5213</td>
<td>0.8734</td>
<td>0.7266</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.8644</td>
<td>0.5867</td>
<td>0.8389</td>
<td>0.8177</td>
</tr>
<tr>
<td>RETAIN</td>
<td>0.8967</td>
<td>0.5808</td>
<td>0.8693</td>
<td>0.8029</td>
</tr>
<tr>
<td>RNN</td>
<td>0.9038</td>
<td>0.6234</td>
<td>0.8636</td>
<td>0.8051</td>
</tr>
<tr>
<td>Proposed Model</td>
<td>0.9428</td>
<td>0.7476</td>
<td>0.9016</td>
<td>0.8424</td>
</tr>
</tbody>
</table>

Deep sequential models (i.e. RETAIN\textsuperscript{23} and RNN\textsuperscript{12}) take the original event sequence as their inputs as described in section 4.1. We implemented our model and neural network based baselines with Keras (https://keras.io).

The event embedding size is set to 32 while the hidden layer size is set to 64. The max number of groups $M$ is set to 32. When training the models, we used Adam\textsuperscript{33} with the mini-batch of 32 samples and the “early stopping” strategy when the performance of validation set drops down.

5.2 Evaluation Metrics

Metrics for binary labels such as accuracy are not suitable for measuring the performance on imbalanced datasets. Therefore, similar to the works\textsuperscript{18,23} we adopt ROC curves (Receiver Operating Characteristic curves) and PRC (Precision-Recall curves) for evaluation metrics. Both of these two curves reflect the overall quality of predicted scores, according to their true labels. To get quantitative measurements, the area under ROC(AUC) and the area under PRC(AUPRC) are utilized.

5.3 Quantitative Results

Table 3 shows the AUC and AUPRC of different models on the death prediction and the ICU admission prediction tasks. From the results shown in Table 3, we can draw the following conclusions:

First, on the whole, deep sequential models(including RETAIN, RNN and the proposed model) outperform non-sequence models(including SVM, LR, and Random Forest) on both tasks, which suggests that temporal information is effective in the outcome prediction tasks.

Second, our model outperforms all the sequential models. For example, on the “ICU admission task”, the proposed model improves AUC by at least 3.4% and AUPRC by at least 3.0% compared other models on all tasks. The improvement verifies our claim that it’s more proper to capture temporal dependencies of clinical event sequences in a hierarchical way.

5.4 Ablation Studies

In this section, we perform ablation studies to examine the effects of our proposed techniques, namely the event attention mechanism, the temporal attention mechanism, and the adaptive segmentation module.

The first study over two attention mechanisms is performed on both tasks. Specifically, we re-train our model by ablating certain components:

- **W/O E-Attn**, where no event attention is performed and a group representation is set as the average of the event embeddings in this group.

- **W/O T-Attn**, where no temporal attention is performed and the sequence representation is set as the final output of the GRU.

Results of attention mechanism ablation studies are represented in Table 4. We can see that both attention mechanisms
Table 4: Ablation study over attention mechanisms

<table>
<thead>
<tr>
<th>Models</th>
<th>Death AUC</th>
<th>Death AUPRC</th>
<th>ICU admission AUC</th>
<th>ICU admission AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>W/O T-Attn</td>
<td>0.9348</td>
<td>0.7181</td>
<td>0.8987</td>
<td>0.8400</td>
</tr>
<tr>
<td>W/O E-Attn</td>
<td>0.9170</td>
<td>0.6404</td>
<td>0.8930</td>
<td>0.8376</td>
</tr>
<tr>
<td>Full Model</td>
<td><strong>0.9428</strong></td>
<td><strong>0.7476</strong></td>
<td><strong>0.9016</strong></td>
<td><strong>0.8424</strong></td>
</tr>
</tbody>
</table>

Figure 2: Ablation study over attention mechanisms. The fix-length segmentation splits the event sequence into groups of equal size (except the last one).

contribute to the strong empirical results of our model represented previously. It is noteworthy that the event attention, one of the important parts of hierarchical representations, plays a more critical role in our model compared to the temporal attention, especially on the death prediction task.

Besides the attention mechanisms, study over the adaptive segmentation is performed on the death prediction task. We re-train our models by replacing the adaptive segmentation module with the fix-length segmentation which splits the original sequence into groups of equal size events (except the last group). Group size of the fix-length segmentation a hyperparameter. Notice that the fix-length segmentation degenerates to no segmentation if group size is set to 1.

Figure 2 shows the AUC and AUPRC of the proposed model where the adaptive segmentation is replaced by the fixed-length segmentation with different group sizes in the death prediction task (the trend in the ICU admission task is similar). We can see that the performance goes down when the group size becomes too small or too large. We infer that if the number is too small, local dependencies of events are modeled as long-term dependencies. And if the number is too large, long-term dependency is lost when the corresponding events are assigned to the same group. Besides, it’s obvious there is a performance gap between the adaptive segmentation and all other segmentation methods, which verifies our claim that the adaptive segmentation can help model long-term dependencies and is suitable for long irregular event sequences.

5.5 Important Events

We analyze the events to which our proposed model pays most attention in prediction. In particular, we use the median of all event attention scores of an event type on a specific task as the importance of the event type on the task.

Top important events on two tasks are listed in Table 5. We can see that even though our model mainly focuses on laboratory tests (such as “Heart Rhythm” and “Blood PH”) on both tasks, the specific events attracting the model on two tasks are different due to their different prediction targets. It is also perhaps surprising that owing to our data-driven approach, our model can select “Family Communication” as an important event type on the death prediction task, which may be ignored by doctors.
Top Events (Sorted by Median of Event Attention Scores)

<table>
<thead>
<tr>
<th>Event</th>
<th>Death Attn Score</th>
<th>Event</th>
<th>ICU Admission Attn Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood Products</td>
<td>0.9965</td>
<td>Blood PH</td>
<td>0.9998</td>
</tr>
<tr>
<td>Radiologic Study: thoracic lumbar sac</td>
<td>0.9896</td>
<td>Vancomycin</td>
<td>0.9995</td>
</tr>
<tr>
<td>NV#2 Waveform Appear: overshoot</td>
<td>0.9713</td>
<td>Hematocrit (35-51)</td>
<td>0.9967</td>
</tr>
<tr>
<td>Heart Rhythm</td>
<td>0.9702</td>
<td>Edimentation rate</td>
<td>0.9885</td>
</tr>
<tr>
<td>Pain Location: periumbilical</td>
<td>0.9668</td>
<td>Daily Weight</td>
<td>0.9850</td>
</tr>
<tr>
<td>Family Communication</td>
<td>0.9523</td>
<td>Bilirubin Total</td>
<td>0.9834</td>
</tr>
</tbody>
</table>

Table 5: Top important events on the death prediction task and the ICU-admission prediction task.

6 Conclusion

In this paper, we proposed a model to learn hierarchical representations of long and irregular clinical event sequences of EHR data for clinical outcome prediction. We validate the performance of our model on real clinical datasets for death and ICU admission prediction tasks. The significant improvements indicated that our model is suitable for irregular timed EHR data and can capture long-term temporal dependencies of clinical event sequences for precise clinical outcome predictions.

Acknowledgements

This paper is partially supported by Beijing Municipal Commission of Science and Technology under Grant No. Z181100008918005, National Key Research and Development Program of China with Grant No. SQ2018AAA010010, and the National Natural Science Foundation of China (NSFC Grant No. 61772039 and No. 91646202).

References


Biomedical Research Cohort Membership Disclosure on Social Media

Yongtai Liu, MS1, Chao Yan, MS1, Zhijun Yin, PhD1, Zhiyu Wan, MS1, Weiyi Xia, PhD1, Murat Kantarcioglu, PhD3, Yevgeniy Vorobeychik, PhD4, Ellen Wright Clayton, MD, JD1, Bradley A. Malin, PhD1,2

1Vanderbilt University, Nashville, TN; 2Vanderbilt University Medical Center, Nashville, TN; 3University of Texas at Dallas, Richardson, Texas; 4Washington University in St. Louis, St. Louis, MO

Abstract

To accelerate medical knowledge discovery, an increasing number of research programs are gathering and sharing data on a large number of participants. Due to the privacy concerns and legal restrictions on data sharing, these programs apply various strategies to mitigate privacy risk. However, the activities of participants and research program sponsors, particularly on social media, might reveal an individual’s membership in a study, making it easier to recognize participants’ records and uncover the information they have yet to disclose. This behavior can jeopardize the privacy of the participants themselves, the reputation of the projects, sponsors, and the research enterprise. To investigate the dangers of self-disclosure behavior, we gathered and analyzed 4,020 tweets, and uncovered over 100 tweets disclosing the individuals’ memberships in over 15 programs. Our investigation showed that self-disclosure on social media can reveal participants’ membership in research cohorts, and such activity might lead to the leakage of a person’s identity, genomic, and other sensitive health information.

Introduction

To accelerate research and improve health care outcomes, various programs are gathering health-related information from individuals to build large cohorts1–3. The primary objective of these programs at the early stage is to collect a wide range of data from their participants, including genomic, phenomic (via surveys and electronic medical records), and demographic information4, 5. The data are then made accessible to researchers to explore hypotheses, study associations, and develop new approaches to manage one’s health6, 7. One common nature of these programs is that they are large and getting larger with respect to the number of participants and the size of collected data. One example of such a program is the Personal Genome Project (PGP), which was launched by Harvard researchers to improve the personalization of medicine4. This program has collected more than 10,000 genomes of participants from a variety of countries8. The 100,000 Genomes Project serves as another example, which has collected the genomes of one hundred thousand British participants to improve research on rare diseases9. And, to investigate how genetic predisposition and environmental exposure contribute to disease development, UK Biobank is now generating whole genome sequencing data on over 500,000 individuals10.

These programs aim to make data widely available, an endeavor that is realized by sharing data with trusted researchers and, at times, with the public2. However, the sharing of individual-level health data raises privacy concerns. This is because participants might consent to making their genome and health data available to researchers (or to the public), but not revealing their identity, which can result in unexpected economic or reputational loss11. As such, the majority of large cohort programs adopt strategies to protect their participants’ identity12, for example, through the application of de-identification routines.

Yet, there are concerns about the degree to which protection can be sufficiently realized in the age of big data. This is because there are various ways in which privacy may be compromised in such systems. For instance, there have been a number of re-identification attacks designed to leverage a wide range of data types13, 14. In 2013, Sweeney and colleagues15 re-identified the names of more than 40% of the PGP participants by linking demographic data (ZIP code, gender, and date of birth) of de-identified records to the voter registration lists. Though these attacks often require a non-trivial amount of time, effort, and money to realize in a manner that would be considered detrimental to a program16, there are several developments that are enhancing the opportunities for penetrating the privacy of individuals in such environments. The first is that participants are increasingly becoming partners in the research environment. The second, and partially an artifact of the first, is that participants are using social platforms to discuss
their experiences in the research domain on a widely accessible scale\textsuperscript{17, 18}. The third is that the research programs themselves may encourage volunteers to tell their stories publicly, with the goal of encouraging people to join the study. Revealing such information makes it evident that the social media sharer is a member of the cohort. This makes it easier for would-be attackers to identify the sharer in the resource. This can be specifically accomplished by using the sharer’s personal information that might be revealed on social media, as well as demographics that might be accessible through information brokers, to link the sharer to their record in the program’s de-identified dataset\textsuperscript{19}. While some individuals may feel comfortable revealing certain information about themselves (e.g., a family history of heart disease), they may not be comfortable revealing their whole genome. As such, this behavior potentially jeopardizes the privacy of the participants themselves, as well as the reputation of the project.

To study the plausibility of an attack, we investigate the frequency of membership disclosure on social media. To do so, we selected a number of research studies from the Database of Genotypes and Phenotypes (dbGap) at NIH and Wikipedia Cohort Study Category\textsuperscript{20}. We then set out to ascertain if any membership disclosure transpired in a popular social media platform, Twitter. To do so, we gathered over 100,000 tweets related to these cohorts and selected approximately 4,000 that contained keywords (e.g., participant, join, volunteer) indicative of potential disclosure. As will be illustrated below, we discovered membership disclosure tweets that revealed the participation of over 100 individuals. We inspected Twitter profiles for these individuals, which indicated demographics, health conditions, and occupations that might be leveraged to link to an individual’s de-identified record. All of the mentioned information provides an opportunity to find the users’ record in the study cohort and uncover additional information that has yet to be revealed in an identified manner, such as the participant’s genome or potentially stigmatizing health information.

This investigation also reveals several patterns. First, we show that membership-related tweets often contain certain types of words (e.g., join, participant, and volunteer). Second, over 80% membership disclosed participants have a non-negative attitude towards the program they are involved in. Sentiment analysis shows that most of these participants are happy to be a part of the cohort, which might be the incentive for some participants to reveal information about themselves. Third, longer lasting and larger cohort studies usually have more membership leakage on Twitter. We note that this is a hypothetical study only and we did not actually re-identify these individuals in the cohorts they claim to be a member of. Nonetheless, our results show that posts on social media can reveal participants’ membership in research cohorts and such activity might lead to the leakage of a person’s identity, genomic and other sensitive health information.

Related Work

The personal health information that has been disclosed on social media has been leveraged to study health-related behaviors\textsuperscript{17, 21, 22}. In spite of the great potential research value, there still exist many concerns regarding the sharing of personal health status or negative health risk behaviors in online environments\textsuperscript{23}. For example, Morgan et al.\textsuperscript{24} showed that one-third of investigated college students reported having posted a picture depicting substance use on social media platforms. Sharing such information will not only trigger privacy concerns about the disclosers themselves (e.g., damage to reputation), but may have the potential to influence other people’s behavior. For instance, it was observed that discussions about prescription abuse over Twitter may aggravate substance abuse\textsuperscript{25, 26}.

Additionally, it should be noted that people share their own information as well as that of other people in online environments. It has been shown that individuals disclose information about a wide range of acquaintances, ranging from family members to friends to high profile persons in the media\textsuperscript{22, 27}. For example, Christofides et al.\textsuperscript{28} illustrated how undergraduate Facebook users posted personal information (e.g., dates of birth and email addresses) in their profiles, but also shared photos of their friends performing potentially sensitive acts (e.g., drinking alcohol at parties).

Our work differs from the aforementioned studies in that we focus on the privacy issues regarding the membership of participants in biomedical research programs on social media. Specifically, we study self-disclosures made by the program participants themselves, as well as investigate the disclosures made by the organizations who own and have responsibility to protect the participants’ data. In doing so, our research contributes to the health information privacy field by highlighting a new type of privacy risk: the cohort membership leakage through social media.
**Cohort Selection**

We selected study cohorts from the database of *Genotypes and Phenotypes (dbGap)* and *Wikipedia cohort studies category*. *dbGaP* was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in humans. It contains 483 biomedical research studies. By contrast, *Wikipedia* provided a convenient list of long lasting cohort studies, such as the *1970 British Cohort Study*. To make our investigation more general, we chose as many different types of studies as possible. The selected cohorts are diverse in three aspects: objective, time, and population.

**Objective.** We selected cohorts to focus both on a specific disease, such as *Type 1 Diabetes Genetics Consortium*, as well as a particular demographic, such as the *Nurses’ Health Study* or gender with in the *Million Women Study*.

**Time.** Cohort studies are not a new phenomenon. Some of the cohorts considered have a long history. For instance, the *Framingham Heart Study* began in 1948. Still, some of the studies are relatively new, for example, the *Qatar BioBank* was launched in 2012. Additionally, we selected studies to have a wide range in duration. Certain longitudinal studies have lasted for decades, while some achieved their objectives in a short period and thus were quite limited in length.

**Population.** The selected cohorts have a varying number of participants. There are multiple cohorts with relatively small sizes, such as the *International HapMap Project*, which collected human genomes from 1,000 participants. By contrast, several cohorts contain hundreds of thousands of participants, such as the *100,000 Genomes Project*.

**Methods**

We partition our search procedure into two steps. Here we provide a high-level overview of the process. The first step is data collection. In this step, we find all of the possible tweets related to the selected cohorts. The second step is data filtering. In this step, we choose a portion of the tweets from step one and manually review these tweets to find the participants of studies. We then perform sentiment and frequency analysis on the tweets that disclose membership in a biomedical research study. The workflow is summarized in Figure 1.

**Figure 1:** The framework for research cohort membership discovery. The processing begins by collecting all tweets that contain the name of cohorts of interest. The tweets are then subject to a membership keyword filter. The remaining tweets are manually reviewed.

**Data Collection.** To collect tweets related to the selected cohorts, we use the names (and abbreviations) of the 77 studies as search keywords and collect all related tweets with a python crawler. By doing so, we obtained 139,529 tweets. Manually reviewing all of the tweets to find those revealing an individual’s participation information would be quite time consuming and error-prone. Since this is a pilot study, and our goal is to demonstrate the possibility of membership disclosure instead of finding all such tweets, we narrowed the scope of our search based on our knowledge to a portion of the tweets that are most likely to contain information about membership disclosures. When we manually reviewed some of the collected tweets, we found that most of the self-disclosed tweets exhibited the following pattern:
“I joined xxx research project today!”,”I am a participant of the xxx program.” or “Now I became a volunteer of the xxx study.”

**Data Filtering.** We filtered the tweets with the following keywords: participant, participate, join, and volunteer, and discarded the remainder of the tweets. It should be recognized that this search method does not guarantee completeness. We lose tweets about disclosure that lack such search terms. For example, “I sent my test sample to xxx project today.” is not caught by the filter. This step yielded 12,698 tweets. For most of the projects, there are fewer than 500 tweets with the keywords of interest. Thus, we manually reviewed all of these tweets to find those that reveal membership disclosure. For cohorts with more than 500 tweets, we randomly select 500 for manual review. Details about the number of tweets collected for each cohort are provided in Table 1. For brevity, we depict the top 50 cohorts that returned the most tweets. Information on all 77 cohorts is available on Github1.

**Sentiment and Frequency Analysis.** The previous step yielded 4,020 tweets. We manually reviewed these tweets, and labeled the tweets containing membership disclosure information. We performed sentiment and frequency analysis on the target tweets posted by project participants. We first removed all the links, hashtags and @ characters from the tweets. We then fed the preprocessed tweets into TextBlob (version 0.15.3) for sentiment analysis. TextBlob is a python package for natural language processing (NLP). For each tweet, TextBlob generates a sentiment score in the range from \([-1, +1]\), where -1 means extremely negative and +1 stands for extremely positive. Next, we partitioned the tweets into words through a process of normalization and tokenization (which partitions a tweet into a set of words), lemmatized (which transforms a word from its original form to its base form; e.g., walks becomes walk) all the words using python NLP package nltk (version 3.3). For the lemmatized words, we removed stop words (e.g., i, ia, in ,the). Since we used cohort names to collect all the tweets, we also dropped all of the words in cohort names, such as “study”, “project”, “health” and “genome”. We then counted the frequency for the remaining words.

**Results**

Table 1 reports the number of tweets collected, filtered and reviewed for 77 selected cohorts. Each of the first six cohorts in Table 1 has more than 10,000 related tweets, which in total accounts for 70% of the total collected tweets. All of the cohorts in the top 25% have over 1,000 tweets. The number of tweets collected from these 19 cohorts accounts for 91.8% of all the tweets. There are 26 cohorts with fewer than 100 related tweets. The distribution of tweets filtered by the selected keywords is roughly the same as the distribution of the total collected tweets. The set of cohorts in the top 6 occupied 87.5% of the filtered tweets and the top 19 cohorts generated 97.6% of filtered tweets. In general, the research programs with larger volume and longer time span have to had more tweets. In particular, programs involving government support often fall into this category.

Among the 4,020 selected tweets, we found 109 that communicated membership disclosure. The results of this investigation are shown in Table 2. These tweets come from 15 of the cohorts (19.5%). They reveal the membership of more than 115 participants. We present some examples of disclosure tweets in Table 3. Notably, 86 of these tweets (78.9%) were posted by cohort participants. In these cases, participants’ leaked either their own or their friends’ membership information when they talked about their experience with some cohort study. This discovery confirms the findings of Yin et al.36 and Mao et al.27, where it was observed that individual’s self-disclosure on social media may reveal other people’s sensitive information. The remaining 23 tweets (21.1%) come from the program’s official account or researcher/organizer of the study. In these cases, the participants’ information was revealed because the program shared a volunteer’s story.

We discuss self-disclosed and program-disclosed tweets separately in the following sections.

**Self-disclosed tweets.** Self-disclosure tweets refer to the tweets posted by cohort participants. These tweets usually have a similar style, such as “I joined/participated in the xxx study” or “I am a participant/volunteer of the xxx program.” Some users wrote an additional sentence to explain why they joined the program or how they feel about it. An analysis of the sentiment of self-disclosed tweets revealed that 71 of the 86 users (82.5%) have a neutral or positive attitude about their participation while 39 of the tweets (45.3%) have a sentiment score greater than 0. Such a positive

---

1https://github.com/yongta1123/Biomedical-Research-Cohort-Membership-Disclosure
Table 1: Number of tweets collected, filtered, and reviewed for 77 cohorts.

<table>
<thead>
<tr>
<th>Study Cohort</th>
<th>All Tweets</th>
<th>Tweets Filtered</th>
<th>Tweets Reviewed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 UK Biobank</td>
<td>24,056</td>
<td>4,265</td>
<td>500</td>
</tr>
<tr>
<td>2 100000 Genomes Project (Genomics England)</td>
<td>22,217</td>
<td>1,735</td>
<td>500</td>
</tr>
<tr>
<td>3 UK 10K</td>
<td>14,999</td>
<td>515</td>
<td>500</td>
</tr>
<tr>
<td>4 LifeLines</td>
<td>14,600</td>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td>5 All of Us Research Program</td>
<td>12,263</td>
<td>4,163</td>
<td>500</td>
</tr>
<tr>
<td>6 National Children Study</td>
<td>10,640</td>
<td>357</td>
<td>357</td>
</tr>
<tr>
<td>7 Human Longevity</td>
<td>4,478</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>8 Qatar Biobank</td>
<td>3,585</td>
<td>296</td>
<td>296</td>
</tr>
<tr>
<td>9 Australian Longitudinal Study on Women's Health (ALSWH)</td>
<td>3,555</td>
<td>136</td>
<td>136</td>
</tr>
<tr>
<td>10 Research Program on Genes, Environment and Health (RPGEH)</td>
<td>2,769</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>11 Personal Genome Project</td>
<td>2,655</td>
<td>436</td>
<td>436</td>
</tr>
<tr>
<td>12 Raine Study</td>
<td>2,538</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>13 Generation Scotland</td>
<td>2,283</td>
<td>106</td>
<td>106</td>
</tr>
<tr>
<td>14 Coronary Artery Risk Development in Young Adults Study</td>
<td>1,616</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>15 Nun Study</td>
<td>1,275</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>16 Millennium Cohort Study</td>
<td>1,195</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>17 Million Women Study</td>
<td>1,182</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>18 Socio-Economic Panel</td>
<td>1,152</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>19 Avon Longitudinal Study of Parents and Children (ALSPAC)</td>
<td>1,005</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>20 Young Lives</td>
<td>919</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>21 LifeGene</td>
<td>850</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>22 Seven Countries Study</td>
<td>833</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>23 Atherosclerosis Risk in Communities</td>
<td>708</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>24 English Longitudinal Study of Ageing</td>
<td>673</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>25 Black Women’s Health Study</td>
<td>619</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>26 International Cancer Genome Consortium</td>
<td>601</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>27 1970 British Cohort Study (BCS70)</td>
<td>600</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>28 Whitehall Study</td>
<td>575</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>29 Nurses’ Health Study</td>
<td>565</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>30 Alameda County Study</td>
<td>353</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>31 Seattle 500 Study</td>
<td>339</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>32 National Child Development Study</td>
<td>299</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>33 Framingham Heart Study</td>
<td>290</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>34 Religious Orders Study</td>
<td>249</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>35 The Irish Longitudinal Study on Ageing</td>
<td>216</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>36 Women’s Interagency HIV Study</td>
<td>184</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>37 Adventist Health Studies</td>
<td>166</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>38 Study of Mathematically Precocious Youth</td>
<td>160</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>39 Newcastle 85+ Study</td>
<td>148</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>40 Great Smoky Mountains Study</td>
<td>129</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>41 International Rare Diseases Research Consortium</td>
<td>128</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>42 UK Households Longitudinal Study</td>
<td>126</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>43 Multicenter AIDS Cohort Study</td>
<td>119</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>44 National Survey of Health &amp; Development</td>
<td>116</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>45 British Birth Cohort Studies</td>
<td>113</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>46 BioBank Japan</td>
<td>103</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>47 MalariaGEN</td>
<td>103</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>48 Taiwan Biobank</td>
<td>102</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>49 COSMOS Cohort Study</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50 Normative Aging Study</td>
<td>100</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Summary | 139,529 | 12,698 | 4,020 |
Table 2: A summary of the cohort and membership coverage from tweets discovered to reveal participation.

<table>
<thead>
<tr>
<th>Study Cohort</th>
<th>Tweets Reviewed</th>
<th>Tweets Disclosed</th>
<th>Self-Disclosed Tweets</th>
<th>Program-Disclosed Tweets</th>
<th>Disclosed Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Personal Genome Project</td>
<td>436</td>
<td>26</td>
<td>26</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>2 100,000 Genomes Project (Genomics England)</td>
<td>500</td>
<td>16</td>
<td>9</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>3 Black Women’s Health Study</td>
<td>22</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>4 Raine Study</td>
<td>72</td>
<td>11</td>
<td>0</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>5 UK Biobank</td>
<td>500</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>6 All of Us Research Program</td>
<td>500</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>7 Qatar Biobank</td>
<td>296</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>8 Nurses’ Health Study</td>
<td>29</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>9 Australian Longitudinal Study on Women’s Health</td>
<td>136</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>10 1970 British Cohort Study (BCS70)</td>
<td>26</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>11 Framingham Heart Study</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>12 Millennium Cohort Study</td>
<td>42</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>13 Million Women study</td>
<td>17</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>14 National Child Development Study</td>
<td>26</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>15 Human Longevity</td>
<td>45</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td><strong>109</strong></td>
<td><strong>86</strong></td>
<td><strong>23</strong></td>
<td><strong>114</strong></td>
<td></td>
</tr>
</tbody>
</table>

attitude shows that most self-disclosed volunteers are happy with the program they participate in and their disclosures on social media express their support or compliment for the program rather than criticism. Words like proud and love often appears in these tweets. Table 4 provides the frequency of the 26 most common words.

At the same time, a small portion of the tweets suggests a negative emotion. For example: “I’ve been a participant for two years, but have not had any feedback.” The distribution of the sentiment score is shown in Figure 2. Self-disclosure tweets usually only reveal the user’s membership; however, at times they may involve their family or close friends. In such cases, one or more of the users’ family members may have a rare disease (e.g., a child who experiences a congenital heart attack) and they joined the research project together to find out why and how to treat it.

Figure 2: Sentiment analysis of 86 self-disclosed tweets. The score ranges from $[-1, 1]$, where -1 means very negative and +1 stands for very positive.

Program-disclosed tweets. At times, the programs post about volunteers’ participation experiences on social media as a way to promote the program and attract the public to join. Most of these tweets reveal a volunteer’s membership often with health information, along with a link to, or a video about, the volunteer’s story. Volunteers talk about why
Table 3: Examples of membership disclosure tweets. We replace the person and cohort names with xxx and rewrite the sentences to mitigate the risk of revealing the program and participants.

<table>
<thead>
<tr>
<th>Type</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Disclosed</td>
<td>1. Proud to be a participant in this: <a href="https://url/abcd">https://url/abcd</a></td>
</tr>
<tr>
<td></td>
<td>2. I like how xxx program never forget my birthday. Thanks @xxx</td>
</tr>
<tr>
<td></td>
<td>3. I joined xxx project because ..., I won’t never share anyone else’s DNA.</td>
</tr>
<tr>
<td></td>
<td>4. I am proud to be a participant in the xxx cohort knowing that I am contributing to a research about health and lifestyle.</td>
</tr>
<tr>
<td></td>
<td>5. I am both a researcher and a participant of the xxx project.</td>
</tr>
<tr>
<td></td>
<td>6. I just volunteered for the xxx project. It was a nice experience, you should try it too!</td>
</tr>
<tr>
<td>Program-Disclosed</td>
<td>1. It’s great to see Mr.xxx and his parents sharing their story about receiving a test result from the xxx research <a href="https://url/abcd">https://url/abcd</a></td>
</tr>
<tr>
<td></td>
<td>2. In this video, meet participant Ms.xxx and her father, xxx, who talked about why taking part is important to them <a href="https://url/abcd">https://url/abcd</a></td>
</tr>
<tr>
<td></td>
<td>3. It’s awesome that @xxx continue to contribute to the Program. Thank you!</td>
</tr>
<tr>
<td></td>
<td>4. XXX, who has heart disease, talks about her participation in xxx study.</td>
</tr>
</tbody>
</table>

Table 4: The most frequent words in 86 self-disclosure tweets.

<table>
<thead>
<tr>
<th>word</th>
<th>count</th>
<th>word</th>
<th>count</th>
<th>word</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant</td>
<td>26</td>
<td>invite</td>
<td>6</td>
<td>would</td>
<td>4</td>
</tr>
<tr>
<td>participate</td>
<td>23</td>
<td>love</td>
<td>5</td>
<td>well</td>
<td>4</td>
</tr>
<tr>
<td>join</td>
<td>20</td>
<td>since</td>
<td>5</td>
<td>remember</td>
<td>4</td>
</tr>
<tr>
<td>get</td>
<td>9</td>
<td>data</td>
<td>5</td>
<td>today</td>
<td>4</td>
</tr>
<tr>
<td>proud</td>
<td>7</td>
<td>one</td>
<td>5</td>
<td>great</td>
<td>4</td>
</tr>
<tr>
<td>interest</td>
<td>6</td>
<td>share</td>
<td>5</td>
<td>learn</td>
<td>4</td>
</tr>
<tr>
<td>years</td>
<td>6</td>
<td>look</td>
<td>4</td>
<td>think</td>
<td>4</td>
</tr>
<tr>
<td>volunteer</td>
<td>6</td>
<td>member</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

they joined the program, as well as what they gained from entering the program. This approach may be useful in attracting people to join the program, but this activity also increases the risk of the volunteer to re-identification.

**Disclosure tweets are more likely to be associated with larger cohorts.** As shown in Table 5, the cohorts with membership disclosure tweets cover more than 10,000 participants. The studies that began more recently tend to have more members active on the Internet, such that they appear to discuss their involvement more often. Some of the tweets posted by participants in long term studies showed that these participants have a stable relationship with the program. These users specifically shared their long term participant experience and feelings about the program. The word “years” appears six times in 32 tweets.

**Tweets can contain search keywords but lack user membership information.** 3,901 of the 4,020 (97.3%) selected tweets do not contain user participant information. Program-related accounts posted most of these tweets and tended to follow one of two patterns. The first is to call for volunteers: “Come and join the xxx research program.” The second is a thank you message to their participants: “xxx participants finished sequencing! Thank you, everyone, for taking part in our research!” On the other hand, tweets posted by users revealed their interest or concern about the program. For example: “I am interested in join the xxx study, but I am worried about my privacy.” In general, it was observed that people are willing to join cohort studies and make their contribution, but a concern of privacy protection is an impediment. For example, 16 tweets talked about the participants’ email address disclosure problem of Personal Genome Project UK.

**Potential Risk of Membership Disclosure.** Based on this analysis, we partitioned the risk of membership disclosure into three types: membership disclosure, identity disclosure and attribute disclosure. Here, we will discuss these privacy threats and illustrate how they relate to the specific population we studied.

613
Table 5: Year launched, number of participants and the number of tweets disclosed for the 15 cohorts.

<table>
<thead>
<tr>
<th>Study Cohort</th>
<th>Disclosing Tweets</th>
<th>Year Launched</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Personal Genome Project</td>
<td>26</td>
<td>2005</td>
<td>10,000</td>
</tr>
<tr>
<td>2 100,000 Genomes Project (Genomics England)</td>
<td>16</td>
<td>2012</td>
<td>100,000</td>
</tr>
<tr>
<td>3 Black Women’s Health Study</td>
<td>12</td>
<td>1995</td>
<td>59,000</td>
</tr>
<tr>
<td>4 Raine Study</td>
<td>11</td>
<td>1989</td>
<td>2,868</td>
</tr>
<tr>
<td>5 UK Biobank</td>
<td>10</td>
<td>2007</td>
<td>500,000</td>
</tr>
<tr>
<td>6 All of Us Research Program</td>
<td>10</td>
<td>2017</td>
<td>20,000</td>
</tr>
<tr>
<td>7 Qatar Biobank</td>
<td>5</td>
<td>2012</td>
<td>20,000</td>
</tr>
<tr>
<td>8 Nurses’ Health Study</td>
<td>4</td>
<td>1976</td>
<td>280,000</td>
</tr>
<tr>
<td>9 Australian Longitudinal Study on Women’s Health</td>
<td>3</td>
<td>1996</td>
<td>57,000</td>
</tr>
<tr>
<td>10 1970 British Cohort Study (BCS70)</td>
<td>3</td>
<td>1970</td>
<td>17,000</td>
</tr>
<tr>
<td>11 Framingham Heart Study</td>
<td>2</td>
<td>1971</td>
<td>14,000</td>
</tr>
<tr>
<td>12 Millennium Cohort Study</td>
<td>2</td>
<td>1991</td>
<td>200,000</td>
</tr>
<tr>
<td>13 Million Women Study</td>
<td>2</td>
<td>1996</td>
<td>1,319,475</td>
</tr>
<tr>
<td>14 National Child Development Study</td>
<td>2</td>
<td>1958</td>
<td>17,415</td>
</tr>
<tr>
<td>15 Human Longevity</td>
<td>1</td>
<td>2013</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Summary 109

The problems induced by membership disclosure are best illustrated with several examples. First, imagine that a volunteer has disclosed his/her membership in some research program. An attacker collects the volunteer’s demographic information (e.g., residential geographical area, gender, and date of birth) from the social network (e.g., the user’s Twitter profile) and links this information to the de-identified participants’ records published by the research program. If a unique linkage to a record transpires, then the attacker has achieved an identity disclosure\(^{19}\). If multiple records are linked to the user, but they share the same (or similar) sensitive attribute value(s), then a successful attribute disclosure attack\(^{37}\) has been perpetrated. Even if their values for the sensitive attribute are different, the attacker can guess the right one with some confidence. By contrast, previous high-profile attacks are limited in that they need to make assumptions about whether a targeted individual is indeed in a dataset. Thus, their claimed attacking powers need to be discounted by the prior probability that a targeted individual has been selected from a broader population\(^{38}\). In our scenario, the attacker is confident that the targeted individual is in the dataset. As a consequence, the discovery of membership significantly increases the likelihood of a successful attack. This attack adds significant power to all the previous attacks, which include the following:

1. Membership Disclosure. As noted earlier, the action of disclosing one’s membership leaks some of the users’ sensitive information. For instance, a project may be disease-specific, such that all of the participants have the same diagnosis. Similarly, some of the users join a study because they, or their relatives, have a rare disease. When they post such information online, their health information is leaked as well.

2. Identity Disclosure. By sharing membership and other personal information over social media, users can be identified. This can be accomplished by collecting self-disclosed users’ personal information from their profile, such as their real name, race, gender, residence, education level, and occupation. To illustrate this issue, we randomly selected ten users and inspected their Twitter profile. It was found that nine out of ten users revealed their real face as their avatar, eight shared their location to a specific city, seven talked about their occupation or education level in their biography, six used their real name as their account name and two users made their date of birth public. With such information on hand, an attacker could find the person through a people search website, such as Intelius.com or InstantCheckMate.com. Moreover, program-disclosed individuals are more readily identifiable because the story shared by programs often contains detailed information about the storyteller. In this case, we learn the volunteer’s personal information from the story, as well as their health information.

3. Attribute Disclosure. Research programs may publish their data to the public or share it with researchers in a de-identified fashion. However, if a malicious attacker has access to the cohort data, along with additional information about the self-disclosing participant (collected from the user’s social media profile), then the attacker can use such information as quasi-identifiers to link to the participant’s record in the cohort database. As mentioned earlier, Sweeney et al. showed that they could identify more than 40% PGP participants using their ZIP code, gender, and date of birth,
and obtain participants’ sensitive information, such as medical conditions and DNA sequence\textsuperscript{15}.

\section*{Discussion and Conclusion}

This investigation illustrates that an individual’s membership in a biomedical research study can be disclosed in social media in several ways. We uncovered tweets that revealed the membership over 100 participants in 15 research programs. Approximately 80\% of the tweets correspond to user self-disclosure, while the remaining correspond to disclosures made by the program organizer. We found that 39 out of 86 (45.3\%) self-disclosed users have a positive attitude towards joined research project. The terms “proud”, “interest”, and “love” were communicated by multiple self-disclosers. The personal information reported in the profiles of the social media users increased the risk of identification, which increases the likelihood that an attacker could link to their record in a de-identified dataset about the cohort, leading to further privacy intrusions, such as the re-identification of genomic information. A program may disclose participants membership when they introduce volunteer and share their story to the public as a way to increase program influence and recruit more participants. These stories may contain personal information and sensitive health information about the volunteer.

Still, there are certain limitations to this work, which pose as next steps for research. First, our search procedure is somewhat ad hoc, such that we failed to detect some tweets about membership disclosure that lack certain words (e.g., participant or volunteer). Second, we studied disclosure behavior only on Twitter, but the same problem may exist in other social platforms, such as Facebook and Instagram. A comprehensive study on additional popular social platforms is needed. Third, the current process requires a final manual review, but it is likely that, with enough instances of disclosure, an automated approach for discovery of such tweets could be developed. At the same time, we believe that if automated approaches can be designed to detect such disclosures, they may also be oriented to assist individuals and program managers to recognize when disclosure is happening inadvertently. It may be that such detection and reflection of the potential risks of such actions may change decisions to reveal such information, and at least lead to more informed decision making.

Mitigating the risk of membership disclosure is not an easy problem to solve. In closing, we wish to offer several possible strategies that may warrant consideration. First, given this threat, research programs could inform participants about the risk of membership disclosure and make it clear that if self-disclosures are made that their privacy may not be guaranteed. At the same time, research programs should inform participants of such threats when asking whether they can share information about participants (e.g., through stories). Alternatively, the program could consider sharing stories without mentioning the volunteers real name or quasi-identifiable information.

\section*{Acknowledgements}

This work was sponsored in part by NIH grants RM1-HG009034, R01-HG006844, and U2COD023196.

\section*{References}

8. Church GM. The personal genome project. Molecular systems biology. 2005;1(1).
12. Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical


Foundations for Studying Clinical Workflow: Development of a Composite Inter-Observer Reliability Assessment for Workflow Time Studies
Marcelo Lopetegui, MD, MS¹, Po-Yin Yin, RN, PhD², Peter Embi, MD, MS³, Philip Payne, PhD²
1 Instituto de Ciencias e Innovación en Medicina, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Chile; 2 Institute for Informatics, Washington University School of Medicine, St. Louis, MO, 3 Regenstrief Institute, Indiana University School of Medicine, Indianapolis, IN, USA;

Abstract
The ability to understand and measure the complexity of clinical workflow provides hospital managers and researchers with the necessary knowledge to assess some of the most critical issues in healthcare. Given the protagonist role of workflow time studies on influencing decision makers, major efforts are being conducted to address existing methodological inconsistencies of the technique. Among major concerns, the lack of a standardized methodology to ensure the reliability of human observers stands as a priority. In this paper, we highlight the limitations of the current Inter-Observer Reliability Assessments, and propose a novel composite score to systematically conduct them. The composite score is composed of a) the overall agreement based on Kappa that evaluates the naming agreement on virtually created one-seconds tasks, providing a global assessment of the agreement over time, b) a naming agreement based on Kappa, requiring an observation pairing approach based on time-overlap, c) a duration agreement based on the concordance correlation coefficient, that provides means to evaluate the correlation concerning tasks duration, d) a timing agreement, based on descriptive statistics of the gaps between timestamps of same-task classes, and e) a sequence agreement based on the Needleman-Wunsch sequence alignment algorithm. We hereby provide a first step towards standardized reliability reporting in workflow time studies. This new composite IORA protocol is intended to empower workflow researchers with a standardized and comprehensive method for validating observers’ reliability and, in turn, the validity of their data and results.

Background and Rationale.

Time-motion studies (TMS) have been widely adopted in biomedicine to study workflow¹. Among the various approaches available, “workflow time studies” provide the most exhaustive approach for understanding clinical workflow¹. In this technique, observers continuously follow a subject for a predefined period of time and record tasks as they occur, producing a sequence of time-stamped tasks².³. This approach allows observers to track unexpected instances of tasks, accounting for task fragmentation, interruptions, and the real-world variability of clinical workflow. Although the introduction of electronic time capture tools has facilitated the recording process by allowing observers to direct their attention on the subjects being studied⁴, there are still concerns of overburdened observers due to the complexity of the data capture process and its effect on data quality.

Like any other method requiring a human data collector, workflow studies are subject to variability and error in the data-capture process. Therefore, well-designed studies often conduct inter-observer reliability assessments (IORA) in order to acknowledge the amount of error introduced to the study due to the inconsistency among data collectors. Although IORA are essential for the overall validity and reliability of a study’s results, IORA are not a common practice in most clinical workflow TMS⁵. IORA are systematically underreported and, when conducted, no specifications on the implementation are provided, raising concerns on the data quality produced by these TMS.

When attempting to conduct IORA in workflow time studies, researchers are faced with three major problems, including a) variability of the observable entity, b) multi-dimensionality of workflow analysis, and c) pairing observations.

a) Variability of the observable entity.

The most-cited method, the Kappa coefficient ($\kappa$), is a landmark in the development of agreement theory and one of the most-used methods to assess reliability in the biomedical sciences⁶. However, unlike most inter-rater reliability assessments (i.e. two radiologists diagnosing x-rays), in TMS the assessment is conducted on an entity that is variable and non-constant over time: clinical workflow can be different for every combination of patient, clinician
and location. Thus, IORA in workflow time studies must be conducted by having two independent observers conduct a synchronous observation, following the same subject at the same time, and without interchanging data or thoughts until the observation is concluded. In view of the nature of clinical workflow, it is impractical to attempt IORA with more than two simultaneous observers due to the disruption of the workflow and an augmented Hawthorne effect caused by space constraints.

b) Multidimensionality of workflow analysis.

Besides the difficulties involved with conducting synchronous IORA sessions, the data schema resulting from these IORA sessions (Figure 1) raises concerns about the suitability of the described statistical methods, and several questions arise. Are workflow researchers analyzing the reliability of identifying the occurrence of relevant tasks? Or are they assessing the agreement in naming those tasks? Or the duration of the tasks? Most studies assess agreement in naming tasks, which results in only a partial assessment of reliability, by focusing only on one dimension of the workflow data and thereby losing the ability to identify other types of errors, like a late start of an observed event.

c) Pairing observations

In a typical implementation of Cohen’s kappa for biomedical research, two observers need to “name” (categorize) a predefined number of “items” (petri dishes, tissue samples, etc.). In contrast, in workflow time studies observers must first recognize the existence of the “items” (detecting that a task is happening), name them, and accurately time-stamp them. This usually results in a different number of observations being recorded by independent observers (one observer identifies multiple tasks while the other observer only identifies one, like the case presented in Figure 1). Which tasks are considered for the paired analysis? And how are they paired?

Problem Statement.

As workflow researchers, we are interested in every aspect of clinical workflow: the sequence (order in which tasks occur), time of occurrence (time when tasks occur), count (number of task occurrences), and duration of tasks required to accomplish an activity or goal. The multi-dimensionality of the data produced by workflow time studies provides the required information for such comprehensive analyses. Thus, each aspect should be taken into account when conducting inter-observer reliability to ensure that we are maintaining the integrity of the data captured.

In a previous review, only 6% of TMS were aware of this issue and, in an initial effort towards conducting a more comprehensive IORA, they attempted to use a combination of two methods\(^7\)\(^8\) (intra-class correlation for time and Kappa for categorization).

Given the current state of the art and the limitations encountered, we hypothesize that a composite score for assessing each dimension would provide a more meaningful and comprehensive methodology to train observers and report reliability in workflow time studies. We aim to develop a comprehensive IORA methodology for workflow time studies, contributing to advancing the standardization of workflow research.

\(^a\) Usually, one of these observers is highly knowledgeable in the environment being studied and participates in the tasks definitions (specifying start and end milestones for each observable task), and is considered to be the gold standard to train less-experienced observers. There are no guidelines or consensus on how to define the gold standard.
Methods.

Research questions involving clinical workflow focus on different aspects of the workflow itself. Questions include: How much time do subjects devote to a task? What do subjects do at any given time? What is the duration of specific instances of tasks? What is the sequence of tasks required to complete an activity? When does a given task occur? Each of these questions requires data precision in different dimensions of the observations. Thus, an IORA should assure the reliability of identifying tasks, including naming and timing those tasks (the agreement on the start time and the duration of the task), and the sequence of tasks being captured.

Given the common use of Kappa in TMS, it could be considered as a sufficient, but limited, approach for assessing IORA in workflow time studies. Kappa provides a meaningful and interpretable score for the dimension being assessed, but can only assess one dimension, which opens a chance for bias and misleading IORA scores, jeopardizing the validity of study results.

To the best of our knowledge, the pairing process of the observations in workflow time studies has never been fully described in the literature. This is not trivial, because a) given that each observer can create a different number of observations during the activity, different pairing permutations are possible; and b) depending on the pairing approach, Kappa is measuring different aspects of the data.

Figure 2 describes two proposed computationally achievable pairing methods to prepare the data for Kappa. The first approach artificially breaks the observation into atomic events of one second of duration. This allows us to calculate an agreement based on the number of seconds that both observers are recording the same task name. We named this approach “proportion-kappa” (PK). In the second approach, researchers must arbitrarily match tasks from both observers. It might seem intuitive to perform the matching based on the order of the tasks; however, it fails with insertions/omission scenarios (when the trainee records more than one task for any given task from the gold standard,
or vice versa). Even if both observers end up with the same number of tasks recorded, it might be due to one of them creating a balanced number of insertions and omissions. To overcome these complex scenarios, the desired task matching is achieved by pairing any given task to the task it overlaps the most (time-wise). This approach seems more natural, since it compares each task as a unit, while being flexible with expected minor time-gaps in tasks’ start times. Once pairs are created, researchers can evaluate the agreement on naming those tasks and/or evaluate the duration of the matched pairs. We called this approach “naming-kappa” (NK).

Although PK measures the agreement on naming artificially-created one-second tasks, it is technically evaluating the agreement on the number of seconds sharing the same task name: the overall length of time observers agree it took to complete a task. On the other hand, although in NK we are matching based on time (overlap), we are technically measuring the agreement on naming.

In both approaches, the Kappa coefficient accurately describes proportion agreement or naming agreement. But using only one dimension IORA, is insufficient to assure data quality. For example, PK could serve as a valid IORA technique if the research question focuses on overall time devoted to specific tasks. However, PK fails if the research question concerns the tasks’ frequency or the average duration of tasks required to complete the activity. Furthermore, PK does not consider the task frequency, and is sensitive to tasks’ duration, which allows bias if there are long idle times during the observation (captured as “others”). Thus, even if observers fail at identifying or timing tasks of interest, as long as they agree in “other” tasks, they might reach a good reliability (although Kappa contributes to reduce this bias compared to percentage agreement alone). On the other hand, NK completely disregards tasks’ duration and is non-sensitive to idle time, but is insufficient if the research question focuses on “how much time subjects devote to a given task”, since the pairing does not consider precise start times or precise durations.

Therefore, we designed a composite score comprising five metrics as described below, the first two of which use the Kappa statistic.

**Metric 1: Proportion-Kappa (PK).**

PK provides an appropriate reliability assessment to evaluate the agreement over time devoted to specific tasks. The result of this agreement provides a corrected estimate of the proportion of time two observers agree on the task name, which ensures data validity for studies focusing on time devoted to specific tasks. The abovementioned limitations of PK are tasks’ frequency and overweighting idle time, which should be compensated by other types of agreement assessments.

**Metric 2: Naming-Kappa (NK).**

NK provides a means to analyze the agreement on naming tasks recorded by the observers, weighing each record equally, and not being influenced by the duration of the task. This balances the concern of overweighting long duration tasks on the agreement introduced by the PK method.

**Metric 3: Duration Concordance Correlation Coefficient (D-CCC).**

The only methods used to assess reliability of tasks’ durations in published TMS are the Bland-Altman and the intra-class correlation coefficient. Although the intra-class correlation coefficient stands as a more suitable approach to assess consistency of the tasks’ durations compared to Pearson’s correlation, like Bland-Altman, it assumes that the true value is unknown, thus only measuring the similarity among the measurements. We propose the use of a more suitable measure of agreement concerning continuous variables for non-exchangeable observers (i.e. having a gold-standard): the concordance correlation coefficient (CCC). In basic terms, the CCC evaluates the degree to which pairs of observations fall on the 45° line through the origin. It is composed by the Pearson correlation coefficient (measuring how far each observation deviates from the best-fit line), and a bias correction factor (measuring how far the best-fit line deviates from the 45° line through the origin). It has been described as a robust test on as few as 10 pairs of data, improving the suitability for short observation periods. The incorporation of this method into our IORA protocol would provide the assessment of data validity for research questions concerning individual tasks’ durations.

In order to implement the D-CCC, paired tasks are also required. Tasks need to maintain duration data, thus the PK second-by-second approach is inappropriate. The “most-overlapping task” pairing approach used in NK preserves tasks’ duration. However, in this case, we are only interested in tasks where observers agree on the name, having in mind that the disagreement in naming is already being penalized in NK. Also, in the insertion or omission scenarios, the NK pairing approach creates multiple pairs of the longest task with the overlapping insertions. However, this metric attempts to provide an assessment of data for research questions concerning individual tasks’ duration: hence each task should be paired only once. Once pairs are created with the NK “most-overlapping” approach, all pairs of tasks with non-matching names are dropped (penalized in NK). Where insertions/omissions occur, the gold
standard task is only paired to the task from the insertion block with the same name. If more than one task with the same name exists in the insertion block, the first instance of the task is matched (see Figure 3). Although we feel inclined to add the duration of all tasks with the same name from the insertion block and compare them against the long overlapping task, we consider that the PK pairing is already doing so. In this case, we are interested in the duration of each instance of a task. Thus, only one task from the insertion block should be paired and evaluated. Although this makes the metric very sensitive for insertion/omission scenarios, it provides an assessment of the impact of the insertion (how early it happened relative to the expected duration of the stopped task).

**Metric 4: Sequence-Needleman-Wunsch (S-NW).**

In order to assess reliability considering the sequence of tasks recorded, regardless of their start time and duration, we exploited the transdisciplinary nature of informatics. Sequence comparison is the most important primitive operation in computational biology, serving as a basis for many other, more complex, manipulations. In the bioinformatics arena, assessing sequence alignment of similar strings is a common need, with existing working solutions that provide optimal alignments for similar DNA strings, usually complemented with a similarity score. This is used to study phylogenetic trees, gene mutations, or to reconstruct full DNA sequences based on a master string (among many other uses). Thus, the arrays of data produced by IORA sessions could be represented as two DNA strings. By removing the time and duration dimensions of the observation (taken care of in other scores), each task would represent a “nucleotide,” thus an observation output could be thought of as a sequence of nucleotides, with insertions left unpaired and penalized, as a nucleotide insertion would be in a DNA alignment.

We propose the use of the Needleman-Wunsch algorithm: one of the most-used global comparison algorithms that relies on dynamic programming to compute the similarity between two sequences. First, a similarity matrix is created, based on arbitrary scores for each match, mismatch, or gap. Then, the optimal alignment is reconstructed based on the similarity matrix, and a score is derived from a pre-computed matrix.

The first row and column of the matrix are initialized with multiples of the gap penalty: they represent the score received if only gaps were created (the alignment if the other string is empty). Then, the other cells are computed based on the three adjacent cells already populated \([gs(i-1),tr(j)], [gs(i-1),tr(j-1)]\) and \([gs(i),tr(j-1)]\). This can be thought of as the three options to match the current “nucleotides-tasks”:

a) Align \(gs(i)\) with \(tr(j-1)\) and pair a space with \(tr(j)\) [\(tr(j)\) being an insertion].

b) Align \(gs(i-1)\) and \(tr(j-1)\) and pair \(gs(i)\) with \(tr(j)\) [\(gs(i),tr(j)\) being either a match or mismatch].

c) Align \(gs(i-1)\) with \(tr(j)\) and pair \(gs(i)\) with a space [\(gs(i)\) being an insertion].

Thus, the score to populate on the cell \([tr(i),gs(j)]\) is given by the equation:

\[
[tr(i),gs(j)] = \max \left\{ \begin{array}{l}
[tr(i),gs(j-1)] - G \\
[tr(i-1),gs(j-1)] + P[tr(i),gs(j)] \\
[tr(i-1),gs(j)] - G
\end{array} \right. 
\]

Where \(G\) corresponds to the gap penalty, and \(P\) corresponds to either the “match” score or the “mismatch” score. Once the similarity matrix is computed, a recursive algorithm computes the optimal alignment. We empirically set the scores of -1 for the gap penalty, 0 for a mismatch, and +3 for a match.

In order to produce a meaningful score to be interpreted by the workflow researcher, we propose to transform linearly the theoretical maximum and minimum alignment scores to 0-1 boundaries. The theoretical maximum corresponds to “P-match” times the number of tasks of the longest string, while the theoretical minimum corresponds to negative difference of tasks between the two strings (only mismatches for the pairs, plus gap penalties for the remainder).

**Metric 5: Timing Agreement (TA).**

The timing agreement represents the degree of synchronization by both observers in recording the start time of a task. Given the lack of formal methods to assess this, we propose to report the median time-gap of paired tasks from the D-CCC assessment, grouped by task: the median time (in seconds) of the difference between start times of paired tasks (positive if the trainee is behind the gold standard, negative if the trainee is ahead). We believe this metric might not play a useful role in reporting agreement as a global score since the interpretation may vary depending on the variance of individual tasks (a median time-gap of 2 might be representing being 10 seconds ahead on task “A”, and late 2 seconds on task “B” and 12 seconds on task “C”). However, this metric might be of great contribution to train observers and evaluate tasks based on the magnitude and the sign of the time-gap. We suggest to report the...
median time-gap by task, accompanied by all the time-gaps for that task. This will provide a quick assessment of any potential difficulties on identifying start times for specific tasks.

In this paper, we synthesized a small data set of an IORA session, and used the proposed composite IORA scores to demonstrate their meaning and interpretation. Implementation of the new methods proposed were based on data set shown in Figure 3.

<table>
<thead>
<tr>
<th>goldStandard</th>
<th>taskName</th>
<th>startTime</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>15</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>20</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>23</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>28</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>46</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>56</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>trainee</th>
<th>taskName</th>
<th>startTime</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>5</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>16</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>21</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>24</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>29</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>46</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>51</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>54</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>56</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3:** Synthesized scenario depicting a brief IORA session with a trained observer (goldStandard) and a trainee.

**Results.**

**Application of the composite IORA protocol.**

**Proportion-Kappa (PK).**

For the synthesized scenario (Figure 3), this score results in a Kappa coefficient of 0.744 and 82% agreement. This demonstrates an overall substantial agreement between the observers, which is particularly relevant if the research question focuses on how much time subjects devote to any given task.

**Naming-Kappa (NK).**

This score results in a Kappa coefficient of 0.718 and 80% agreement, demonstrating a substantial agreement on naming tasks between the observers, which is of interest if the research question relates to “what do subjects do at any given time?”.

**Duration Concordance Correlation Coefficient (D-CCC).**

In the synthesized scenario, the trainee performed well, confusing task D with C, and creating an insertion of B while A was described by the gold standard. The “most overlapping” pairing approach produced 10 pairs of tasks. Following the proposed modifications, 3 pairs are dropped: D(5)-C(5) since observers failed at naming them, and A(10)-B(3) and A(10)-A(2) since they correspond to the left-over after pairing the insertion (see Figure 4).

Thus, using Lin’s$^9$ concordance correlation coefficient in STATA (concor command) results in 0.906. Unlike Kappa, there is no specification in the literature on a descriptive scale for the degree of agreement achieved with Lin’s CCC, only existing a lower bound for an acceptable CCC of 0.75 (not universally accepted)$^{13}$. 

**Figure 4:** The “most overlapping” pairing approach produced 10 pairs of tasks, displayed at the right with the duration of the tasks between parentheses. Then, three pairs are dropped in order to implement the D-CCC: one non-matching pair and two pairs of the insertion block.
Sequen

Figure 5 shows the similarity matrix corresponding to the synthesized scenario, based on the strings gs= O-A-D-B-A-C-A-B and tr= O-A-C-B-A-C-A-B, for the gold standard and the trainee observer, respectively. In the synthesized scenario, the computed maximum score is 19 (Figure 5). The theoretical maximum corresponds to 3x10, while the theoretical minimum corresponds to -2. Thus, the linear transformed final score for the current example would be 0.656 ([19+2]/ [30+2]). This is a quantified interpretation for a visual inspection: 7 out of the 10 tasks were correctly matched, one mismatch, and two insertions.


In the synthesized scenario, accuracy for timestamping each task was: 0 seconds for A (0;0;1), 0.5 seconds for B (0;1), and 1 second for C (1), based on the pairing achieved using the D-CCC pairing shown in Figure 4.

Expanding the impact of a meaningful IORA: contributing to an optimized training feedback.

The composite IORA metric intends to empower researchers with a means of ensuring reliability and validity of the data in a quantitative, comparable, and systematic manner. However, there is room for a further contribution: besides informing the researchers “how good” or “how bad” observers are doing based on the metrics calculated, we thought to include visualizations that would help better understand how the scores are calculated and provide an answer to “what” and “when” they disagree. For example, both Kappa metrics, the contingency table highlights discrepancies (any cell not in the diagonal). Also, a side-to-side visualization of both observations permits a global qualitative IORA, which, although not useful for reporting purposes, might be very helpful in training the observers, highlighting times and tasks that produce discrepancies. Similar to the clinical workflow analysis tool (CWAt) and the sample report of Mache’s tool, we developed a graphic representation for an IORA session (Figure 6).

Finally, a simple scatter plot of the durations assessed by the D-CCC helps identify where the timing discrepancies are occurring. The scatter plot depicts the concordance between the duration of each task captured by the gold standard and the trainee. Everything below the plot’s purple diagonal represents a task captured by the trainee having a logged duration that is shorter than the gold standard; anything above represents a task captured by the trainee having a logged duration that is longer than the gold standard (Figure 6).
Discussion.

As previously described, studying workflow goes beyond simply calculating duration of events. In order to fully understand workflow, analyses focus on what happens, when, and for how long, and usually analyze the order in which those tasks occur (the sequence required to complete an activity). Having evidenced the limitations of using a single statistical test to ensure reliability of the observation, and given the lack of a single exhaustive test to accomplish comprehensive IORA, we postulate that using multiple instances of existing tests, one for each dimension, might provide IORA in a more meaningful manner.

Given our experience designing and conducting time-motion studies, achieving scores consistently high, like the one in our synthesized scenario, is not common, and reliability efforts shouldn’t pursue optimal scores in every dimension. Although we recommend assessing every dimension, agreement goals should be directed to the specific research question under study.

Besides the contribution as a reporting tool, our proposed IORA stands as a major contribution to the standardization of the observers’ training, which is usually only reported as total hours of training. With our IORA

---

**Figure 6:** Composite score example. Interpretation for this report: Overall, both observers have an almost perfect PK agreement (relevant if data intends to represent the overall time devoted to any given task). Strong agreement for NK. D-CCC: satisfactory (not directly interpretable, most paired tasks on 45-degree line). Sequence agreement: satisfactory (2 insertions out of 20 recordings). Training recommendation based on report: Review specifications on when “treatment” (blue) ends [avoid trainee early stop, that created the insertion]. Review when “education” (green) begins [improve timing agreement]
approach, researchers have powerful means to direct efforts on observers’ training, by both visually examining details of the observation and by evaluating agreements trends over time in the training period. Thus, instead of reporting hours of observers’ training alone, they will be able to report training until X agreement is achieved in Y dimension (the one of more relevance for the research question).

These metrics were implemented and calculated to produce meaningful and interpretable results, however the D-CCC and S-NW haven’t been used before, and thus interpretations and extrapolation of the scores aren’t straightforward. Although range and directionality of the scores are logical, further intensive testing and validation studies are required to define meaningful cut points. This will allow us to study IORA scores over long periods of observations (assessing observers’ fatigue), study IORA at different points of the study, not just the training portion (assessing observers’ drift), and many other interesting related issues such as the effect of the initial number of tasks on IORA.

Limitations.

The PK, NK, and D-CCC are statistical estimates, and confidence intervals for the estimates are dependent on the sample size. For the PK, a large number of pairs is usually created (number of seconds in the observation), and narrow confidence intervals can be achieved. However, for the NK, the number of pairs corresponds to the number of tasks recorded and, for the D-CCC, to the number of tasks recorded minus the discrepant tasks. Thus, when the number of tasks recorded is small, the NK and D-CCC are bounded by large confidence intervals, jeopardizing the significance and meaningfulness of the scores obtained.

The S-NW and the TA interpretation are not affected by sample size. The S-NW provides a normalized arithmetic score, interpretable independent of the sample size (number of tasks recorded in the observation): insertions are more heavily penalized in shorter observations and vice-versa. Likewise, the TA is a composite descriptive statistic (the median and quartiles), interpretable regardless of the sample size.

We recommend that, if only a small number of tasks are recorded during an observation, individual scores should be carefully considered. We believe the PK might be the most representative score to report in those cases.

Potential improvements to our proposed IORA include the use of weighted Kappa to calculate the agreements. Based on an ontology of commonly used tasks in healthcare, mismatches could be penalized differently depending on how similar or different the two concepts are.

Despite the limitations of a percentage agreement, we believe that, since observers are trained before the IORA session and little guessing is likely to exist, providing the Kappa statistics accompanied by the percentage agreements might be of aid in interpreting the agreement16.

Conclusion.

Confirming that the establishment of IORA protocols and guidelines are a priority in validating continuous observation TMS, we evidenced the limitations of using a single metric IORA, and limitations related to the data manipulation required to implement it. We confirmed that assessing inter-observer reliability with Kappa, although partially useful, misrepresents and overestimates the real agreement, since it only focuses on one dimension of a multi-dimensional data set.

We proposed a composite IORA protocol, including a set of methods to assess each relevant dimension in workflow time studies: a proportion agreement, naming agreement, duration agreement, sequence agreement, and timing agreement. We demonstrated our proposed IORA methodology in a synthesized scenario.

We hereby provided a first step towards a standardized reliability reporting in workflow time studies. This new composite IORA protocol is intended to empower workflow researchers with a standardized and comprehensive method for validating observers’ reliability and, in turn, the validity and representativeness of the data collected.

References.
Comprehensive Application to Support Data Capture for Time Motion Studies. AMIA Annu Symp Proc. 2012;
Bootstrapping Adversarial Learning of Biomedical Ontology Alignments

Ramon M. Maldonado, MS¹, Sanda M. Harabagiu, PhD¹
¹University of Texas at Dallas, Richardson, TX, U.S.A.

Abstract

Learning how to automatically align biomedical ontologies has been a long-standing goal, given their ever-growing content and the many applications that rely on them. Because the knowledge graphs underlying biomedical ontologies enable neural learning techniques to acquire knowledge embeddings as representations of these ontologies, neural learning can also consider ontology alignments. In this paper, we present the Knowledge-graph Alignment & Embedding Generative Adversarial Network (KAEGAN) which learns (a) to represent the relational knowledge from two distinct biomedical ontologies in the form of knowledge embeddings and (b) to use them for ontology alignment, by also relying on the ontology semantics. KAEGAN is a Generative Adversarial Network trained using bootstrapping to iteratively improve the learned alignments. Experimental results show promise, demonstrating that jointly learning ontology alignment and knowledge representation improves upon learning either in isolation.

Introduction

Over the past two decades, the biomedical research community has increased its efforts to produce ontologies encoding biomedical knowledge, justified by the steady increase in biological and biomedical research and the growth of data that is being collected in all areas of biology and medicine. Not only is the number of ontologies increasing and their size growing, but their relevance in biomedical research is also rising as (a) they contribute to the interpretation of the biomedical data and (b) enable complex inference from their encoding. The BioPortal* of the National Center for Biomedical Ontology (NCBO) is the most comprehensive repository of biomedical ontologies in the world (as of this writing it includes 760 ontologies, with over 9 million classes and almost 40 million indexed records). Many of the ontologies available from the BioPortal became widely used resources, as evidenced by thousands of publications citing them. However, many of the ontologies exhibit overlapping information, which imposes the task of ontology alignment, i.e., finding mappings or correspondences between concepts and relations in different ontologies¹.

Biomedical ontology alignment has been performed using various methods, including rule-based and statistical methods²-⁴. More recently, neural methods for incorporating contextual and background information have been applied to ontology alignment as well. Wang et al.⁵ propose a neural model that learns to align concepts from different ontologies using external definitions and context information. Kolyvakis et al.⁶ approach the problem from the perspective of representation learning, aligning concepts using the word embeddings of their descriptions and automatically generated description paraphrases. However, while both approaches determine concept matches using learned embeddings of the concepts in each ontology, neither approach informs their concept embeddings with the relational knowledge contained in the ontologies. Biomedical ontologies encode knowledge by capturing ontological and biomedical expertise in the form of the knowledge graph of structured relations between medical concepts specified in an ontology. The task of distilling the structured knowledge from a knowledge graph into concept and relation embeddings is called knowledge graph embedding.

Recent work⁷-⁹ in learning knowledge graph embeddings has shown that ontological knowledge can be successfully expressed as real valued vectors associated with (i) each concept and (ii) each relation type available in the knowledge graph. The resulting learned knowledge embeddings can be used in a wide range of applications¹⁰ including determining the plausibility of any possible relation between arbitrary concepts in the ontology. In our previous work¹¹, we have shown how knowledge embeddings learned from the Universal Medical Language System (UMLS)¹² improve the performance of a neural prediction model when these knowledge embeddings are used to enrich the model’s representation of medical concepts. Furthermore, knowledge graph embedding methods have also been recently applied to general-domain and cross-lingual knowledge graph alignment with promising results¹³,¹⁴. Intrigued by these recent findings, we sought to extend knowledge-embedding-based alignment techniques to the biomedical domain in this work. For this purpose, we devised the Knowledge-graph Alignment & Embedding Generative Adversarial Network.

*http://bioportal.bioontology.org
KAEGAN method for jointly learning knowledge graph alignment and embedding. KAEGAN embeds the knowledge graphs associated with two distinct biomedical ontologies into the same semantic embedding space such that the learned knowledge embeddings can also be used for ontology alignment. KAEGAN uses a novel adversarial learning framework to learn alignment-oriented knowledge embeddings. More specifically, KAEGAN uses multiple generators to model interactions inside and across two knowledge graphs in order to embed them into the same semantic space, ensuring that aligned concepts will have similar embeddings. The ontology alignment is bootstrapped by iteratively predicting new alignments informed by the most similar alignment-oriented knowledge embeddings. We show that by jointly learning knowledge graph alignments and knowledge embeddings for any pair of biomedical ontologies, we improve the results of learning either knowledge embeddings or knowledge alignments in isolation. Therefore, the main novelty of KAEGAN arises from its adversarial framework which enables joint learning of knowledge alignments while learning how to discriminate valid relations in each knowledge graph, both with competitive results.

Data

In this work we focused on the knowledge embedding representation and alignment of three ontologies available from BioPortal: SNOMED Clinical Terms, the National Cancer Institute (NCI) Thesaurus, and the Foundational Model of Anatomy (FMA). SNOMED CT, the largest of the three ontologies, encodes medical terms used in clinical documentation and reporting, consisting of 349,548 medical concepts as of January 31, 2019. The NCI Thesaurus is an ontology providing a reference terminology for 66,724 cancer and cancer-related medical concepts. The Foundational Model of Anatomy (FMA) is an ontology with the stated goal of representing the phenotypic structure of the human body, encoding 78,989 concepts. Each of these ontologies is distinct and was designed independently. Consequently, a subset of the concepts in each ontology also have representations in the other two ontologies. As each of these ontologies has been integrated into the Unified Medical Language System (UMLS), we used UMLS as a reference alignment for evaluating the quality of the learned knowledge alignments. We were inspired by the Ontology Alignment Evaluation Initiative’s (OAEI) Large BioMed Track, which in fact uses these three ontologies and the reference alignment provided by the UMLS to evaluate ontology alignment systems in a yearly competition. The OAEI Large BioMed track is an evaluation for unsupervised systems and does not concern itself with distilling ontological knowledge into embeddings. As such, we have adapted the alignments provided by the OAEI to the supervised knowledge graph alignment problem.

SNOMED Clinical Terms, the NCI Thesaurus and FMA proved to be ideal ontologies for our work since they not only have a widely accepted alignment evaluation, but each ontology contains rich relational knowledge that can be used to inform the learning of knowledge embeddings, enabling KAEGAN to jointly learn knowledge graph alignment and embedding.

Methods

To learn knowledge embeddings informing ontology alignments we have designed the Knowledge-graph Alignment & Embedding Generative Adversarial Network (KAEGAN), which learns (1) how to embed two distinct biomedical ontologies into the same semantic space such that pairs of concepts from different ontologies can be aligned based on their similarity, while also (2) encoding the semantics of both ontologies in the same, shared embedding space. In KAEGAN, knowledge alignments are learned using the structure of the ontologies, i.e. relying on (a) the relations spanning concepts and (b) the attributes of concepts. KAEGAN also leverages a Generative Adversarial framework to learn high-quality alignment-oriented knowledge embeddings, which in turn are used to iteratively bootstrap the learning of ontology alignments. To describe the functionality of KAEGAN, we first present the way in which alignments between ontologies are learned and then we detail the way in which KAEGAN learns alignment-oriented embeddings.

Learning Ontology Alignments. Formally, let $C_X$ and $C_Y$ denote the set of concepts encoded in ontologies $X$ and $Y$, respectively. An alignment $A = \{(c_x, c_y) \in C_X \times C_Y | c_x \equiv c_y\}$ is a set of pairs of concepts from $X$ and $Y$ that

---

1http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/
represent equivalent concepts across the ontologies. For example, the concept “Arterial structure” from SNOMED CT represents the same concept as “Artery” in the NCI Thesaurus (concept C0003842 in the UMLS), so there would be a pair (Arterial_structure, Artery) in A. It should be noted that $c_1 \equiv c_2$ cannot hold if $c_1$ and $c_2$ are from the same ontology, so a concept $c_x$ from X can be aligned with at most one concept $c_y$ from Y, and vice-versa. We consider a subset of $A, A'$ as training data. This allows us to model alignment as a classification problem where the probability, $q(c_y|c_x)$, of a concept $c_y$ being aligned with a concept $c_x$ is a function of the similarity of their concept embeddings:

$$q(c_y|c_x) = \text{softmax}(\text{sim}(c_x, c_y)) = \frac{e^{\text{sim}(c_x, c_y)}}{\sum_{j \in Y} e^{\text{sim}(c_x, c_j)}}$$ (1)

$$\text{sim}(c_x, c_y) = \frac{\bar{v}(c_x) \cdot \bar{v}(c_y)}{\|\bar{v}(c_x)\|_2 \|\bar{v}(c_y)\|_2}$$ (2)

where $\bar{v}(c_x)$ represents the concept embedding for $c_x$, whereas $\text{sim}(\cdot, \cdot)$ is computed by the cosine similarity.

Inspired by Trivedi et al.\textsuperscript{14}, we adopt a contextualized concept embedding method that is particularly well suited to knowledge graph alignment. LinkNBed\textsuperscript{14} contextualizes concept embeddings using (1) concept attributes and (2) the neighborhood of nearby concepts in the knowledge graph. In a biomedical ontology, concept attributes encode auxiliary information via attribute triples of the form $(c, t, v)$ where $c$ represents a concept, $t$ represents an attribute type, and $v$ represents the attribute value, usually a string. For example in the National Cancer Institute Thesaurus\textsuperscript{19}, the concept ε=“Eicosapentaenoic Acid” has the attribute $t=$“definition” value $v=$“A class of polyunsaturated fatty acids with 20 carbons and 5 double bonds”. To further contextualize each concept embedding, nearby concept embeddings are aggregated using random walks in the knowledge graph. Formally, the embedding $\bar{v}(c)$ for the concept $c$ in KAEGAN is calculated as:

$$\bar{v}(c) = \sigma(v_0(c) + W_n N_e(c) + W_a A_e(c))$$ (3)

where $\sigma$ is the sigmoid function, $v_0(c) \in \mathbb{R}^d$ is an initial $d$-dimension embedding of $c$; $N_e(c)$ is the aggregate neighborhood context embedding of $c$; and $A_e(c)$ is the aggregate attribute embedding of $c$, while $W_n, W_a \in \mathbb{R}^{d \times d}$ are weight matrices. More specifically:

- The aggregate neighborhood context embedding $N_e(c)$ is computed by averaging the initial embedding vectors $v_0(c_i)$ for each concept $c_i$ in the neighborhood of $c$. The neighborhood of a concept $c$ is approximated as the set of concepts other than $c$ encountered on $k$ random walks of length $l$ executed when starting at $c$.

- The aggregate attribute embedding $A_e(c)$ is computed by max-pooling over the attribute embeddings of each attribute of $c$.

An attribute embedding $a$ is calculated by passing an attribute type embedding $a_t$ and an attribute value embedding $a_v$ through a fully connected sigmoid layer: $a = \sigma(W_t a_t + W_v a_v)$, where $W_t, W_v \in \mathbb{R}^{d \times d}$ are weight matrices. Attribute type embeddings are learned from scratch for each attribute and attribute value embeddings are learned using PARAGRAPH2VEC\textsuperscript{20} as in Trivedi et al.\textsuperscript{14}.

Using the concept embeddings to calculate $q(c_y|c_x) \forall(c_y, c_x) \in C_Y \times C_X$, we can measure the quality of a predicted alignment according to $q$ using cross entropy:

$$- \sum_{x \in X} \sum_{y \in Y} 1_{(c_x \equiv c_y)} \log q(c_y|c_x)$$ (4)

However, Equation 4 will only measure how well the model captures similarity between concepts aligned in the training data, which represents a small subset of $C_X \times C_Y$. Inspired by Sun et al.\textsuperscript{13}, we extended Equation 4 to incorporate uncertainty for unlabeled alignments, using the function $\phi(c_x, c_y)$ in place of the indicator function of Equation 4:

$$\phi(c_x, c_y) = \begin{cases} 1 & \text{if } c_x \text{ is aligned in the training data} \\ \frac{1}{N_{\text{unl}}} & \text{if } c_x \text{ is unlabeled} \end{cases}$$ (5)

where $N_{\text{unl}}$ is the number of currently unaligned concepts from $Y$. $\frac{1}{N_{\text{unl}}}$ represents a uniform distribution over the possible alignment candidates for $c_x$ and serves to bias the system against erroneous alignments. Using $\phi$ in the
cross-entropy calculation of Equation 4 in place of the indicator function, we obtain the Alignment Classification loss, $L_C$:

$$
L_C = - \sum_{x \in X} \sum_{y \in Y} \phi(c_x, c_y) \log q(c_y|c_x)
$$

By minimizing $L_C$, KAEGAN learns the probability alignment function, $q$, which enables the maximum likelihood alignment between $X$ and $Y$ to be found. As in Sun et al.\textsuperscript{13}, we produce an alignment between $X$ and $Y$ given an alignment probability function, $q$, by solving the max-weighted matching problem on the bipartite graph whose nodes are concepts from $X$ and $Y$ with edges $\langle c_x, c_y \rangle$ denoting alignment weighted by $q(c_y|c_x)$. We only consider alignment between each concept pair $\langle c_x, c_y \rangle$ for which $q(c_x, c_y)$ is above a certain threshold. It should be noted that this represents a max-weighted, one-to-one matching between subsets of $X$ and $Y$ with maximum total likelihood according to $q$.

Moreover, Sun et al.\textsuperscript{13} have shown that such alignments can be iteratively refined through bootstrapping by adding newly predicted alignments to the training data at each iteration and altering the predictions if a more likely alignment emerges. Specifically, at iteration $i$, given a training alignment $A'_i$ (where $A'_0 = A'$), KAEGAN learns alignment-oriented knowledge embeddings for each concept in $C_X \cup C_Y$ using the adversarial method described in the next section, after which the alignments are learned and new concept matches are added to $A'_{i+1}$. If a more likely matching emerges for a particular concept according to $q$, the less likely matching is simply replaced. KAEGAN terminates when no new concept matchings are added to the alignment.

**Adversarial Learning of Alignment-Oriented Knowledge Embeddings.** In order to learn alignment-oriented knowledge embeddings KAEGAN uses a Generative Adversarial Network\textsuperscript{16} (GAN) composed of a Discriminator and four Generators illustrated in Figure 1. In the traditional GAN framework, two competing networks - a Discriminator and a Generator - are pitted against one another in a zero-sum game. The Generator learns to map from a latent space to a data distribution of interest, while the Discriminator learns to distinguish candidates produced by the generator from the true data distribution. Recently, GANs have been applied to the problem of Knowledge Graph Embedding\textsuperscript{8,11}, showing promising results. In this work, we extend the GAN framework for learning knowledge graph embedding\textsuperscript{8} to learn alignment-oriented knowledge embeddings. In KAEGAN, the Discriminator learns alignment-oriented embeddings that are used to discriminate between valid and invalid relation triples from both biomedical

---

**Figure 1:** Adversarial Learning of Knowledge Graph Embeddings for Biomedical Ontology Alignment.
ontologies\footnote{A relation triple $\langle c_1, r, c_2 \rangle$ is defined by a relation encoded between two concepts $c_1$ and $c_2$ from the same biomedical ontology.} while ensuring that any aligned concepts from the pair of ontologies have similar embeddings, given a probability of alignment between the concepts, defined as $q$. The relation triple $\tau = \langle c_1, r, c_2 \rangle$ is considered to be valid if the concept $c_1$ is related to the concept $c_2$ by the relation $r$ in either of the knowledge graphs $X$ or $Y$ and is said to be invalid otherwise. For example, in the NCI Thesaurus, the concept “Artery” is encoded as a type of “Blood Vessel” represented by the relation triple $\langle \text{Artery, IS\_A, Blood\_Vessel} \rangle$. The Discriminator should be able to discriminate between such a valid relation and an invalid relation such as $\langle \text{Artery, IS\_A, Opiate} \rangle$. Observing that the majority of randomly constructed invalid relation triples can be trivially recognized as invalid, the four Generators of KAEGAN learn to produce more plausible – yet still invalid – relation triples in an attempt to fool the Discriminator, e.g. $\langle \text{Artery, IS\_A, Blood\_capillary} \rangle$. In order to facilitate alignment, the Discriminator is trained to evaluate the plausibility of both intra-graph relation triples produced by Generators $G_X$ and $G_Y$ and inter-graph relation triples produced by Generators $G_{XY}$ and $G_{YX}$.

Consider two relation triples $\tau_x = \langle c_1^x, r^x, c_2^x \rangle$ and $\tau_y = \langle c_1^y, r^y, c_2^y \rangle$ from the knowledge graphs represented by ontologies $X$ and $Y$, respectively shown in Figure 1. Attempting to fool the Discriminator, Generator $G_X$ uses $\tau_x$ to generate an invalid triple $\tau'_x = \langle c_1^x, r'^x, c_2^x \rangle$ by swapping out either $c_1^x$ or $c_2^x$ with another concept from $X$ such that $\tau'_x$, while invalid, is more plausible than a randomly sampled triple (e.g. $\tau_x = \langle \text{Artery, IS\_A, Blood\_Vessel} \rangle$ and $\tau'_y = \langle \text{Artery, IS\_A, Blood\_capillary} \rangle$). Generator $G_Y$ does the same for triple $\tau_y$, generating the invalid triple $\tau'_y$ using $Y$. The Discriminator’s job is then to determine which of $\tau_x$ and $\tau'_x$ and which of $\tau_y$ and $\tau'_y$ is more plausible using the loss function, $L_H$. Generators $G_{XY}$ and $G_{YX}$ operate similarly, however they generate invalid triples by sampling concepts across the knowledge graphs corresponding to the pair of biomedical ontologies. Specifically, $G_{XY}$ generates the invalid triple $\tau'_{xy}$ by first relying on the valid triple $\tau_{xy}$ and replacing either $c_1^y$ or $c_2^y$ with its aligned concept from $Y$, $c_1^x$ or $c_2^x$, based on an alignment $A'$, available from the training data used for learning ontology alignments. Then $G_{XY}$ samples a concept from $Y$ to replace $c_1^y$ or $c_2^y$ to produce $\tau'_{xy}$. For example, consider the triple $\tau_x = \langle \text{Artery, IS\_A, Blood\_Vessel} \rangle$ from FMA. Blood_Vessel could be replaced with the concept Human_Blood_Vessel from NCI to create $\tau_{xy} = \langle \text{Artery, IS\_A, Human\_Blood\_Vessel} \rangle$ and the generator might sample the semantically related concept Angiogram to produce $\tau'_{xy} = \langle \text{Artery, IS\_A, Angiogram} \rangle$. $G_{YX}$ operates in the same way to create $\tau_{yx}$ and $\tau'_{yx}$ from $\tau_y$.

In order to discriminate valid relation triples from invalid relation triples, the Discriminator is trained to measure the plausibility of a triple $\tau = \langle c_1, r, c_2 \rangle$ where $c_1, c_2 \in C_X \cup C_Y$ are concepts and $r$ is a relation from either $X$ or $Y$. KAEGAN models relation triple plausibility scores using translational knowledge graph embedding. Pioneered by TransE\cite{bordes2013translating}, translational knowledge graph embedding methods model relation triple plausibility as a translation on the embedding space. TransE models a relation triple $\tau = \langle c_1, r, c_2 \rangle$ by first projecting $c_1, c_2$, and $r$ onto a shared embedding space to produce embeddings $\vec{v}(c_1), \vec{v}(c_2)$, and $\vec{v}(r) \in \mathbb{R}^d$, where $d$ is the dimensionality of the embedding space. The vector $\vec{v}(r)$ is used as a translation vector from $\vec{v}(c_1)$ such that $\vec{v}(c_2) \approx \vec{v}(c_1) + \vec{v}(r)$ if the triple $\tau = \langle c_1, r, c_2 \rangle$ is a valid relation triple. In this way, TransE models the plausibility of a relation triple using the distance between the destination concept embedding and the source concept embedding translated by the relation embedding: $f_D(\tau) = \| \vec{v}(c_1) + \vec{v}(r) - \vec{v}(c_2) \|_{L2}$. The concept embeddings $\vec{v}(c_1), \vec{v}(c_2)$ are computed using Equation 3 and the relation embedding $\vec{v}(r)$ is calculated via simple embedding lookup.

In order to learn the plausibility of the relations encoded in an ontology, KAEGAN minimizes the following marginal loss function:

$$L_R = \sum_{\tau \in X \cup Y} \max(0, f_D(\tau) - \gamma_1) + \max(0, \gamma_2 - f_D(\tau'))$$

where $\gamma_1, \gamma_2$ are margin hyperparameters and $f_D(\tau')$ is the score of an invalid triple generated by either $G_X$ or $G_Y$ using $\tau$. While the triple $\tau$ represents semantic knowledge that we wish to encode in our knowledge embeddings,
\( \tau' \) represents erroneous knowledge that KAEGAN should be implausible. The two margin parameters \( \gamma_1, \gamma_2 > 0 \) where \( \gamma_2 > \gamma_1 \) enforce the property that plausible triples have lower score than negative triples since \( f(c_1', r, c_2') - f(c_1, r, c_2) \geq \gamma_2 - \gamma_1 > 0 \). By modeling the margin with two parameters and setting \( \gamma_1 \) to a small positive value, we also enforce the property that plausible triples have low absolute scores, which has shown to be effective for alignment-oriented embedding\(^{13,21}\).

The corrupted triple \( \tau' \) is sampled by either Generator \( G_X \) or Generator \( G_Y \) depending on whether the original triple \( \tau \) was encoded in ontology \( X \) or in ontology \( Y \). Given a triple \( \tau_x \in X \), Generator \( G_X \) generates the corrupted triple \( \tau'_x \) by producing a probability distribution \( P_X(\tau'_x | \tau_x) \) and sampling from that distribution. The distribution \( P_X \) is calculated using Generator \( G_X \)'s own scoring function \( f_{G_X} \):

\[
P_X(\tau'_x | \tau_x) = \frac{e^{f_{G_X}(\tau')}}{\sum_{\tau^* \in Neg^X} e^{f_{G_X}(\tau^*)}}
\]

where \( Neg^X \) is a set of invalid relation triples derived from \( \tau_x = (c_1^x, r^x, c_2^x) \) with either \( c_1^x \) or \( c_2^x \) replaced with another concept from ontology \( X \) such that the resulting relation triple does not occur in the knowledge graph \( X \). KAEGAN uses the DistMult\(^8\) scoring function, defined as \( f_{G_X} = (\vec{v}(c_1) \odot \vec{v}(r)) \cdot \vec{v}(c_2) \), where \( \odot \) is element-wise multiplication and \( \cdot \) is the dot product. DistMult has been shown to excel in probabilistic systems that learn knowledge embedding models\(^11\). Intuitively, \( P_X \) is used to sample the most plausible, yet still incorrect triples in an attempt to fool the Discriminator. Generator \( G_Y \) samples from the distribution \( P_Y \) calculated using the scoring function \( f_{G_Y} \) defined similarly.

For alignment-oriented embeddings, the scoring functions \( f_D, \) (of the Discriminator) and \( f_G \) (of the generators) should have the replacement property that if \( c_1^x \in X \) has an aligned concept \( c_1^y \in Y \), \( f(c_1^x, r^x, c_2^x) \approx f(c_1^y, r^x, c_2^y) \) when \( f \) is either \( f_D \) or \( f_G \). That is to say, if a concept from a plausible triple of ontology \( X \) is replaced with an aligned concept from ontology \( Y \), the new triple should remain plausible under both scoring functions. The same property should hold if \( c_2^x \) is replaced with its aligned concept or if a concept from a triple from ontology \( Y \) is replaced with an aligned concept from ontology \( X \). The replacement property of \( f_D \) and \( f_G \) ensures that relation semantics are preserved across knowledge graphs. In order to capture this property, we minimized the cross-graph marginal loss, defined as:

\[
L_A = \sum_{\tau_x \in X^A} \max(0, f_D(\tau_{xy}) - \gamma_1) + \max(0, \gamma_2 - f_D(\tau'_{xy})) \\
+ \sum_{\tau_x \in Y^A} \max(0, f_D(\tau_{yx}) - \gamma_1) + \max(0, \gamma_2 - f_D(\tau'_{yx}))
\]

where \( X^A, Y^A \) represent the subsets of triples in the knowledge graphs of the ontologies \( X \) and \( Y \) that involve concepts aligned across ontologies; \( \tau_{xy} \) represents a triple \( \tau_x = (c_1^x, r^x, c_2^x) \in X \) with either \( c_1^x \) or \( c_2^x \) replaced with its aligned concept from \( Y \), and \( \tau'_{xy} \) represents a triple \( (c_1'^x, r^x, c_2'^x) \in X \) with either \( c_1'^x \) or \( c_2'^x \) replaced with an incorrect concept from \( Y \) that does not represent a true alignment, sampled by \( G_{XY} \). \( \tau_{yx} \) and \( \tau'_{yx} \) are defined similarly.

As in previous work\(^9,11\), because the Generators involve a sampling step, they were trained via Policy Gradient\(^22\) where each Generator was trained to maximize the reward it receives from the Discriminator, in the form of the Discriminator’s score of the generated relation triple, \( f_D(\tau'_x), f_D(\tau'_y), f_D(\tau'_{xy}), \) and \( f_D(\tau'_{yx}) \).

### Results

The alignment-oriented knowledge embeddings learned by KAEGAN were evaluated in terms of their ability to (1) produce an alignment between two biomedical knowledge graphs, and (2) model the semantics of the two encoded knowledge graphs. For the evaluations, we used the three ontologies from the OAEI large-bio track described in the data section: the National Cancer Institute Thesaurus\(^6\), the Foundational Model of Anatomy\(^7\), and a subset of SNOMED CT\(^5\). For the alignment task, we randomly sampled

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Training</th>
<th>Validation</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNOMED-NCI</td>
<td>3,442</td>
<td>1,721</td>
<td>12,047</td>
</tr>
<tr>
<td>SNOMED-FMA</td>
<td>1,205</td>
<td>602</td>
<td>4,219</td>
</tr>
<tr>
<td>NCI-FMA</td>
<td>537</td>
<td>269</td>
<td>1,880</td>
</tr>
</tbody>
</table>

**Table 1:** Alignment data distribution.
| Model          | P     | R     | F₁    | H@10  | MRR   | P     | R     | F₁    | H@10  | MRR   |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| KA-Classifier | 0.8410| 0.5734| 0.6819| 0.8507| 0.7735| 0.8204| 0.6058| 0.6969| 0.8425| 0.7395|       |
| KAEGAN-No Class | 0.5508| 0.3199| 0.5480| 0.5847| 0.4973| 0.5210| 0.4323| 0.4725| 0.583| 0.4794|       |
| KA-Discriminator | 0.8372| 0.5910| 0.8629| 0.7632|        | 0.8317| 0.6239| 0.7103| 0.8497| 0.7383|       |
| KAEGAN-No Context | 0.8402| 0.5880| 0.8662| 0.7572|        | 0.8237| 0.6159| 0.7048| 0.8417| 0.7433|       |
| KAEGAN-No Attr | 0.8573| 0.6125| 0.7145| 0.8896| 0.7842| 0.8466| 0.6310| 0.7231| 0.8668| 0.7655|       |
| KAEGAN        | 0.8410| 0.5734| 0.6819| 0.8507| 0.7735|       |       |       |       |       |       |

**AML²**

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>H@10</th>
<th>MRR</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>H@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>KA-Classifier</td>
<td>0.6752</td>
<td>0.5575</td>
<td>0.6107</td>
<td>0.6905</td>
<td>0.6155</td>
<td>0.6720</td>
<td>0.5540</td>
<td>0.6067</td>
<td>0.6835</td>
<td>0.5933</td>
</tr>
<tr>
<td>KAEGAN-No Class</td>
<td>0.8265</td>
<td>0.6565</td>
<td>0.7318</td>
<td>0.8403</td>
<td>0.7694</td>
<td>0.8318</td>
<td>0.6238</td>
<td>0.7125</td>
<td>0.8524</td>
<td>0.7570</td>
</tr>
<tr>
<td>KA-Discriminator</td>
<td>0.5365</td>
<td>0.4735</td>
<td>0.5030</td>
<td>0.5413</td>
<td>0.5190</td>
<td>0.5361</td>
<td>0.4326</td>
<td>0.4778</td>
<td>0.5714</td>
<td>0.5004</td>
</tr>
<tr>
<td>KAEGAN-No Context</td>
<td>0.5338</td>
<td>0.4679</td>
<td>0.4987</td>
<td>0.5435</td>
<td>0.5190</td>
<td>0.5380</td>
<td>0.4279</td>
<td>0.4756</td>
<td>0.5716</td>
<td>0.5054</td>
</tr>
<tr>
<td>KAEGAN-No Attr</td>
<td>0.8175</td>
<td>0.6665</td>
<td>0.7343</td>
<td>0.8323</td>
<td>0.7724</td>
<td>0.8271</td>
<td>0.6235</td>
<td>0.7103</td>
<td>0.8427</td>
<td>0.7576</td>
</tr>
<tr>
<td>KAEGAN</td>
<td>0.8419</td>
<td>0.6815</td>
<td>0.7533</td>
<td>0.8493</td>
<td>0.7922</td>
<td>0.8486</td>
<td>0.6417</td>
<td>0.7303</td>
<td>0.8686</td>
<td>0.7806</td>
</tr>
</tbody>
</table>

**AML²**

Table 2: Evaluation of biomedical ontology alignment using alignment-oriented knowledge embeddings.

20% of the aligned concepts of each pair of ontologies to create our initial training Alignment, $A_0$, reserving 10% for validation, leaving the remaining 70% as test data, summarized in Table 1. For both tasks, we evaluate KAEGAN against the following alternate configurations of KAEGAN to explore the efficacy of the methods used therein:

1. **KA-Classifier** is a neural classifier that learns a similarity function $q(\cdot)$ which minimizes the objective in Equation 6, but does not use the relational information of the two knowledge graphs being aligned. KA-Classifier is meant to evaluate the effect of relational information on embedding-based alignment. KA-Classifier does not use either loss from Equations 8 or 10 and therefore can not be trained in an adversarial fashion.

2. **KAEGAN-No Class** is a configuration of the KAEGAN system that does not use the classification loss, $L_C$. KAEGAN-No Class learns to embed and align concepts using only the relational information from the two knowledge graphs as represented by $L_R$ and $L_A$. This configuration serves to evaluate the effect of $L_C$ on both alignment and embedding.

3. **KA-Discriminator** is a model comprised only of the Discriminator of KAEGAN, meant to evaluate the effect of the GAN training regime. KA-Discriminator is trained using standard negative sampling instead of the Generators.

4. **KAEGAN-No Context** is a configuration of KAEGAN that uses simple embedding lookup for $\tilde{v}(\cdot)$ in Equation 3, eschewing context information from attributes and concept neighborhoods.

5. **KAEGAN-No Attr** is a configuration of KAEGAN that does not use attribute information, leaving $W_a A_c(c)$ out of Equation 3.

6. **KAEGAN-No Neigh** is a configuration of KAEGAN that does not use neighborhood information, leaving $W_n N_c(c)$ out of Equation 3.

Following the Ontology Alignment Evaluation Initiative³, ontology alignment is evaluated using Precision (P), Recall (R), and $F_1$ as well as Hits@10 and Mean Reciprocal Rank as in Sun et al.¹³. Hits@10 reports the percentage of concepts $c_x$ for which the true aligned concept $\hat{c}_y$ is one of the top-10 most likely matches according to $q(c_y|c_x)$. Mean reciprocal rank reports the average reciprocal rank of the correct aligned concept among all possible concepts, ranked by $q$. 

633
The ontology alignment results are presented in Table 2. We compare the different configurations of KAEGAN to the top performing system from the OAEI Large BioMed track, AgreementMakerLight (AML), a hand-engineered expert system which leverages background information found outside the ontologies being aligned. While not at the level of AML, clearly, the full KAEGAN model out-performs the alternate configurations for each matching. Interestingly, the worst performing configurations were the KAEGAN-No_Context and KAEGAN-No_Attr models, illustrating the importance of attribute information for ontology alignment. Moreover, the strong performance of the KA-Classifier model and the relatively poor performance of the KAEGAN-No_Class model demonstrate the efficacy of modeling alignment matching explicitly through the alignment classification loss, $L_C$.

To determine the ability of the learned knowledge embeddings to model the semantics of the embedded knowledge graphs, we evaluate the embeddings based on their plausibility and completeness following the evaluation procedure outlined in previous work. Plausibility evaluation is cast as a link prediction problem called Relation Triple Classification (RTC). RTC is performed using a plausibility function defined as $\rho = -f_D$, the negation of the scoring function of the KAEGAN’s Discriminator. Given a relation triple $\tau = (c_1, r, c_2)$, we use $\rho(\tau)$ to classify $\tau$ as either a valid or invalid relation triple depending on if $\rho(\tau) > \omega$ for some threshold $\omega$. To perform RTC, we split the relation triples from each knowledge graph into 85%/5%/10% train/validation/test splits. For each triple in each validation and test set, we create an invalid triple by replacing either the source or destination concept and add the invalid triple to the respective set. We then use the trained model to score each triple from the validation set and select the optimal value of $\omega$ for each model. We evaluate RTC on the test sets using Precision (RTC-P) and Recall (RTC-R) as in previous work. RTC-P can be thought of as measuring plausibility while RTC-R measures completeness. To further quantify plausibility, we use three standard knowledge graph embedding metrics: (1) Pairwise Plausibility Accuracy (PPA); (2) Hits@10 (H@10); and (3) Mean reciprocal rank. PPA measures how often a model can select the valid triple from a pair of triples $(\tau, \tau')$ where $\tau'$ is a corrupted version of $\tau$. H@10 and MRR measure how well the model can use plausibility, $\rho$, to rank unseen valid triples above invalid triples. To compute H@10 and MRR, we first create a set of rankings, one for each triple in the test set. Each ranking is created by replacing either the source or destination concept and add the invalid triple to the respective set. We then use the trained model to score each triple from the validation set and select the optimal value of $\omega$ for each model. We evaluate RTC on the test sets using Precision (RTC-P) and Recall (RTC-R) as in previous work.

The results presented in Table 3 indicate that KAEGAN is able to model the semantics of distinct knowledge graphs by embedding them in the same semantic space facilitated by modeling their alignment. The KA-Classifier model does not perform knowledge graph embedding, so it is not evaluated along with the other configurations. Likewise, we do not evaluate KAEGAN against general-purpose, non-alignment-oriented embedding models (e.g. Trans-X, SePLi²², or TATEC²⁴) in order to focus on the impact of modeling alignment on knowledge graph embedding. Although it should be noted that the general KAEGAN learning framework can accommodate any general purpose embedding paradigm,
such as the ones listed above, by replacing Equation 3. The improvement of KAEGAN over KAEGAN-No_Class shows that using alignment classification improves relational learning. Likewise, the improvement of KAEGAN over KAEGAN-Discriminator demonstrates the efficacy of the GAN learning framework. It should be noted that, while the removal of concept context in the form of attributes and neighborhoods has a marked effect on performance, the decrease is minimal compared to its effect the alignment task.

**Discussion**

In general, the results show that KAEGAN is able to embed distinct knowledge graphs into the same semantic space such that the learned embeddings can be used for both ontology alignment and representation learning. The ontology alignment experiments presented in Table 2 elucidate several interesting phenomena including the importance of attributes in ontology alignment. Since attributes are useful in qualifying the concepts - making them distinct from the other concepts in the ontology - it is not unexpected that ignoring them results in the large performance drops seen by KAEGAN-No_Context and KAEGAN-No_Attr. The name and description attributes are particularly helpful, since many concepts have similar names or descriptions across ontologies. However, the performance of KAEGAN-No_C demonstrates the efficacy of leveraging the structured knowledge in the form of relation triples for alignment. One possible reason for this is that concepts with differing names and descriptions from different ontologies may share similar relations with more well-defined concepts, causing the model to discover their alignment. Another interesting trend we can glean from the alignment experiments is the small improvement of Hits@10 over Precision, a fundamentally more difficult evaluation. Precision is akin to ‘Hits@1’, but when we relax the evaluation to the top 10 we only see a 2% increase on average. This finding indicates that when the model finds an alignment it is generally correct, but when the model is not able to determine alignment correctly, the probability it learns for the correct alignment is especially low.

While the alignment results show promise, they do not surpass the current state-of-the-art on the OAEI Large BioMed track, represented by AgreementMakerLight (AML)\(^4\). AML is hand-engineered for aligning large biomedical ontologies, leveraging outside ontologies\(^{25-27}\) as intermediaries to facilitate matching. We believe KAEGAN can be improved in the same way – by jointly modeling the alignment more than two ontologies. We plan to address extend KAEGAN to an arbitrary number of ontologies future work. Likewise, noting that AML makes use of string matching, we believe KAEGAN can be improved by incorporating character-level information about a concept’s name and other attributes.

Unlike AML, KAEGAN jointly performs knowledge embedding as well as alignment. The results presented in Table 3 indicate that biomedical knowledge embeddings learned from distinct graphs are improved by jointly modeling the alignment between the graphs. We believe the primary reason for this is that by modeling both graphs in the same semantic space, each graph can inform the other - filling possible knowledge gaps and providing additional context for both overlapping concepts and the rest of the graphs. Guo et al.\(^{28}\) have shown that semantically smooth knowledge embeddings can be achieved by imposing constraints on the learned embeddings. The alignment classification loss \(L_C\) plays a similar role in KAEGAN constraining aligned embeddings to be similar to one another, imposing its own type of smoothing which is shown to improve results.

**Conclusion**

In this paper, we presented the **Knowledge-graph Alignment & Embedding Generative Adversarial Network (KAEGAN)**. KAEGAN learns to jointly embed distinct biomedical ontologies into the same semantic embedding space such that the resulting embeddings can be used for knowledge graph alignment. KAEGAN leverages relational knowledge encoded in the knowledge graph of an ontology as well as attributes of the medical concepts to learn alignment-oriented embeddings. In addition to ontology alignment, the learned embeddings can be used to model the semantics of the encoded knowledge graphs. Results indicate that learning to align and embed the knowledge graphs jointly improves upon learning the alignment and the embedding separately.

**References**


Barriers, Facilitators, and Potential Solutions to Advancing Interoperable Clinical Decision Support: Multi-Stakeholder Consensus Recommendations for the Opioid Use Case

Laura Haak Marcial, PhD1, Barry Blumenfeld, MD2, Christopher Harle, PhD3, Xia Jing, MD, PhD4, Michelle S Keller, MPH5, Victor Lee, MD6, Zhen Lin, PhD, MS, RN7, Anna Dover, PharmD, BCPS8, Amanda M. Midboe, PhD9, Shafa Al-Showk, MPH, CHES10 Victoria Bradley, DNP, RN, CPHIMS, FHMSS11, James Breen, PharmD8, Michael Fadden, MD12, Edwin Lomotan, MD10, Luis Marco-Ruiz, PhD13, Reem Mohamed, PharmD8, Patrick O'Connor, MD, MA, MPH14, Douglas Rosendale, DO, FACOS, FACS15, Harry Solomon16, Kensaku Kawamoto, MD, PhD17

1RTI International, Rockville, MD; 2RTI International, Waltham, MA; 3Regenstrief Institute, Indianapolis, IN; 4Ohio University, Athens, Ohio; 5Cedars-Sinai Medical Center, Los Angeles, California; 6Clinical Architecture; 7Zhen Research, Houston, TX; 8First Databank, Inc.; 9VA Palo Alto Health Care System, Stanford University, San Francisco, CA; 10Agency for Healthcare Research and Quality, Rockville, MD; 11Lexington, KY; 12Cerner Corporation, Kansas City, MO; 13Norwegian Centre for E-Health Research, Tromsø, Norway; 14Health Partners Institute, Bloomington, MN; 15Deloitte Consulting, Arlington, VA; 16Chicago, IL; 17University of Utah, Salt Lake City, UT

Abstract

With the advent of interoperability standards such as FHIR, SMART, CDS Hooks, and CQL, interoperable clinical decision support (CDS) holds great promise for improving healthcare. In 2018, the Agency for Healthcare Research and Quality (AHRQ)-sponsored Patient-Centered CDS Learning Network (PCCDS LN) chartered a Technical Framework Working Group (TechFWG) to identify barriers, facilitators, and potential solutions for interoperable CDS, with a specific focus on addressing the opioid epidemic. Through an open, multi-stakeholder process that engaged 54 representatives from healthcare, industry, and academia, the TechFWG identified barriers in 6 categories: regulatory environment, data integration, scalability, business case, effective and useful CDS, and care planning and coordination. Facilitators and key recommendations were also identified for overcoming these barriers. The key insights were also extrapolated to CDS-facilitated care improvement outside of the specific opioid use case. If applied broadly, the recommendations should help advance the availability and impact of interoperable CDS delivered at scale.

Introduction

In 2018, while overall opioid prescribing appears to be on a downward trend, misuse of prescription opioids continues to remain fairly stable and opioid related deaths continue to rise.1 Opioids are the second most commonly abused substance in the United States.2 Of the significant efforts to curb this epidemic, one is the development and dissemination of opioid prescribing recommendations for chronic pain from the Centers for Disease Control and Prevention (CDC).3 Part of the challenge in implementing change, however, is applying recommended practices in
routine clinical care. Clinical decision support (CDS) integrated with the electronic health record (EHR) and combined with staff- and patient- oriented strategies, can change clinician behavior and increase adherence to clinical guidelines, including prescribing, but CDS implementation, adoption, and use remains challenging.

While the United States healthcare system has widely adopted EHR technology, and while that adoption generally includes CDS, the overall development and utilization of effective CDS remains somewhat limited. Emerging standards such as the Health Level 7 International (HL7) Fast Healthcare Interoperability Resources (FHIR), CDS Hooks, and Clinical Quality Language (CQL) specifications are becoming increasingly well adopted by EHR vendors, providing a critical link to fully integrated CDS. Requiring support for application programming interfaces (APIs) for enhancing data portability and third-party application integration is also an important facilitator of CDS integration. Interoperable CDS may potentially finally unlock the benefits of converting to electronic records, particularly when patients are engaged through patient-centered CDS.

Despite the promise of widespread care improvement facilitated by interoperable CDS, the reality is that this promise remains a vision for healthcare in general. There are emerging efforts to address this challenge in multiple domains including the opioid epidemic. However, much more needs to be learned on how best to leverage interoperable CDS to improve care widely.

In 2015, the Patient-Centered CDS Learning Network (PCCDS LN) was established to improve collective learning of patient-centered CDS by the Agency for Healthcare Research and Quality (AHRQ). In 2018, the PCCDS LN focused on the opioid epidemic as an important care domain for identifying how best to leverage patient-centered CDS to improve patient care. To do so, two Working Groups were chartered: an Opioid Action Plan Working Group (OAPWG) and a Technical Framework Working Group (TechFWG). The primary goal of the OAPWG was to develop a plan for patient-centered CDS applied broadly and effectively to improve pain management and opioid use. One work product of the OAPWG was the development of highly desirable, consensus patient-centered CDS-enabled future state vision scenarios for patient-centered CDS-enabled pain management/opioid use.

Led by K. Kawamoto and L. Marcial, the primary goal of the TechFWG was to review this future state vision through the lens of CDS implementation and identify the current state, barriers, what resources are available, and recommendations for moving forward. This Task Force was also charged with identifying implications that can be moved forward beyond opioids. Efforts of the TechFWG focused on applying patient-centered CDS in health care delivery and the needs of the healthcare team including the prescribing clinician, with some thinking around the patient perspective. Key considerations of the opioid use case context included:

1. The need for complete and accurate understanding of past, current, and potentially most beneficial treatments in making pain treatment decisions.
2. The ability to quickly and accurately assess opioid-related risks (e.g., OUD, overdose, misuse/abuse).
3. The call to better integrate, synthesize, and interpret information in PDMP reports and other urine drug screen results.
4. To more efficiently and effectively identify trends in relevant patient outcomes over time (e.g., pain, physical function, co-morbid mental health).
Here, we describe the findings from the PCCDS LN TechFWG which encompass barriers, facilitators, implications and recommended actions for advancing patient-centered CDS to address the opioid epidemic.

**Methods**

*Subject matter expert (SME) recruitment.* Participation in the TechFWG by SMEs was voluntary. Calls for participation were sent via email to the PCCDS LN mailing list and to the list serves of the HL7 CDS Work Group, HL7 Clinical Quality Improvement (CQI) Work Group, American Medical Informatics Association (AMIA) CDS Work Group, and AMIA Implementation Work Group. Interested parties were asked to complete an online form with basic demographic information, availability, and interest in serving on the TechFWG.

*TechFWG goals and charter.* The TechFWG was chartered to address four goals: 1) Facilitate shared decision-making tools delivered via CDS, 2) Discuss opportunities and barriers presented by technologies and standards such as SMART on FHIR and CDS Hooks, 3) Identify gaps in needed standards and their adoption, and 4) Produce a whitepaper and/or manuscript on barriers and facilitators.

*TechFWG deliberations.* A total of 8 web conference calls were held biweekly between June 11, 2018 and October 1, 2018. All participants were asked to contribute both during and between meetings and to volunteer to present work of their own to the broader group on CDS-related initiatives. Final report-out and vetting of the recommendations occurred during the PCCDS LN Annual Conference on October 15, 2018 in both a general session as a briefing and a breakout session to solicit additional input.

*TechFWG composition.* The TechFWG included stakeholders from various disciplines. The TechFWG members who contributed in at least 4 of the 8 calls or otherwise made substantial contributions to the final deliverables were invited to participate as co-authors of this manuscript.

*Deliverables.* The deliverables of the TechFWG, and the focus of its activities, consisted of 1) the identification of barriers, facilitators, and required actions to enable effective patient-centered CDS for pain management as envisioned by the OAPWG. Identified barriers were grouped into categories through group consensus, 2) Insights on these barriers, facilitators, and required actions for effective patient-centered CDS in general, extrapolated from these issues identified in the pain management use case, and 3) A summary of the findings and recommendations and its dissemination. This manuscript presents results of the TechFWG deliberations in terms of these deliverables.

**Results**

*Membership and participation:* A total of 54 subject matter experts (SMEs) volunteered to serve on the TechFWG. Participation averaged about 20 members per biweekly call, and 20 members contributed to at least 4 of the 8 calls or otherwise made substantial contributions to the TechFWG deliverables including this manuscript. Of these 20 members, 7 were affiliated with academic institutions, 4 with consulting firms, 3 with government, 3 with research organizations, 2 with EHR vendors, and 1 with a health system (many participants had overlapping roles).

*Barriers, Facilitators, and Needed Actions for Advancing PCCDS for Pain Management:* The TechFWG members identified key barriers in six categories: (1) regulatory environment, (2) data integration, (3) scalability, (4) business case, (5) effective and useful CDS, and (6) care planning and coordination. These are detailed in Table 1 and summarized below (additional details available...
from the PCCDS Learning Network). Based on the findings from the specific use case of pain management, key generalizable insights were extrapolated. These generalizable insights and recommendations are summarized in Table 2.

**Regulatory environment:** Key identified barriers in the regulatory environment include differences in state-level regulations on the allowed use of prescription drug monitoring program (PDMP) data and allowed prescribing of controlled substances. Important identified facilitators in this area included initiatives that connect PDMPs across states as well as state regulations requiring e-prescribing of controlled substances. *Needed Actions:* In terms of recommended actions, the TechFWG recommended unifying regulations according to the evidence base on a national level, encouraging voluntary coordination at the state level, and, at the local level, defining and sharing cohesive evidence-based guidelines that incorporate federal and state regulations.

**Data integration:** As a key challenge, methods for capturing source data, ensuring its quality and availability, specifying the data format, facilitating patient matching, reading relevant data from and writing data into records, and visualizing relevant data can be restrictive or limited. While these data issues are large, pervasive, and perhaps daunting barriers, the TechFWG felt that a facilitator included the fact that opioids are an important national concern with a defined scope, thereby enabling a directed conversation on how specific needed data can be aggregated and analyzed effectively. *Needed Actions:* In terms of recommendations, the TechFWG recommended improving data integration by defining data needs and interoperability requirements tied directly to their clinical application. At the federal level, this might include funding the development of a set of data needs for specific decision makers, a set of consolidated requirements specifications for interoperability, and research on visualization best practices to support clinicians and patient decision making. For providers and patients, we can facilitate improved data quality by giving patients access to update their own data. Vendors need opportunities to more fully test and verify architecture and standards for semantic interoperability in an iterative fashion.

**Scalability:** Assuming the needed data are available, tools to scale patient-centered CDS are still somewhat nascent and often require the additional step of localization. There is substantial promise in emerging repositories to support the availability of CDS artifacts. In addition, relevant standards and support of those standards continues to mature. *Needed Actions:* To facilitate improvements in scalability, the TechFWG identified the need for a national authoritative body that agrees on and promotes human readable and computer-interpretable guidelines and funds research on best implementation strategies for desired CDS. Professional societies continue to develop and provide clinical endorsement of guidelines, while also encouraging widespread adoption of an appropriate technical framework to support interoperable CDS based on these guidelines. Vendors can also provide support in the form of approaches to better centrally manage trusted CDS.

**Establishing a clear business case and providing incentives:** The return on investment for CDS has still not been clearly established and the incentives of today are not yet well aligned with the needs of patients and providers. The advent of value-based care initiatives may help provide some of the needed incentives to advance CDS. *Needed Actions:* Broadly, we must establish and promote the business case for CDS using incentives across stakeholder groups to improve patient care. CMS and other payers could help by incentivizing the achievement of quality goals and best practices through the use of CDS while also funding research to identify and mitigate unintended consequences. Federal and state entities can support and further incentivize the implementation of CDS solutions. The CDS ecosystem can leverage all of its stakeholders to find ways to improve the performance and reduce the cost of CDS.
### Table 1. Key barriers, facilitators, and needed actions for enabling effective PCCDS for pain management

<table>
<thead>
<tr>
<th>Category</th>
<th>Barriers</th>
<th>Facilitators</th>
<th>Needed Actions</th>
</tr>
</thead>
</table>
| **Regulatory environment**    | • There are state-specific regulations on controlled substances, but many states prohibit incorporating the actual PDMP source data into the EHR and CDS.  

  • State-by-state regulations for prescribing limits and/or regulations restrict the ability to prescribe opioid antagonists and there may be limits or incentives for those that can be written electronically which subsequently limits associated CDS.  

  • Federal regulations and recommendations may not align with mechanisms to facilitate CDS implementation.  

  | • Initiatives that link resources up help, e.g., PMP InterConnect, a product from the National Association of Board of Pharmacy that connects PDMPs from 44 states.  

  • In some states there is mandated e-prescribing of controlled substances, e.g., New York.  

  • The Promoting Interoperability program includes 2 opioid e-prescribing measures: query of PDMP (optional in CY 2019 and required in CY 2020) and verification of an opioid treatment agreement (optional for CY 2019 and CY 2020).  

  | • Federal and state lawmakers can:  

  o support the development of unified national-level regulations and encourage voluntary coordination among state medical boards. NOTE: In general, there is a relative lack of evidence underlying regulations and policy.  

  o enable the sharing of source data, and mandate or encourage e-prescribing for controlled substances to further enable point-of-care CDS  

  o create policies and regulations based on evidence and support collection of evidence where lacking |
| **Data integration**          | • It is difficult to access all relevant data across multiple health systems, and health IT systems over time (e.g., medications, supplements, lab tests, imaging, referrals, care plan/controlled substance agreements, functional status, diagnoses and side effects).  

  • Source data may have poor data quality, are temporally limited, and are difficult to both patient-level match and to de-duplicate and may not be in a structured, standard form that facilitates matching, reconciliation, and de-duplication  

  • Existing data structures were created without semantic interoperability as the goal, bi-directional information exchange may not be well supported, e.g., no clear way for controlled substance agreements to be pushed out by a prescriber to other prescribers in a state.  

  • A lack of consensus on a ‘data interoperability architecture’ for universal use inhibits achieving greater overall interoperability.  

  • Visualizations to support interpretation are often limited and scales are subjective, e.g., pain scale.  

  • SMART and other interoperable apps/services may be important in this ecosystem for opioid use  

  | • Opioid use and pain management, a national focus, may serve as a catalyst for CDS.  

  • There is a need to collect well-defined data for aggregation and summarization. The OpenNotes movement could be leveraged for data cleaning and reconciliation. Some patients and health systems are already engaged in Open Notes and it could be used for patients to review and clean data. Natural Language Processing can also be used to mine free text, e.g., for after-the-fact conversion to structured data or for facilitating real-time data entry.  

  • A dashboard for relevant data, e.g., pulling up first-line treatments that have already been tried.  

  • There are some semantic interoperability standards available and adopted for relevant data that could facilitate data integration and summarization:  

  • There are some semantic interoperability approaches/efforts that could be applied to help (e.g., Argonaut project, efforts at HIT Advisory Committee (HITAC) efforts to define priority healthcare uses and needed standards).  

  • Population health management tools  

  | • Concrete interoperability requirements need to be defined for this space. Need to identify achievable goals with associated clinical needs. This applies equally to CDS/knowledge interoperability (see also Scalability).  

  • Federal government could fund the development of a consolidated requirements specification for interoperability.  

  • Data quality and de-duplication issues are still persistent. Need to enable patients to help in this de-duplication and data reconciliation process. Patient involvement will help ensure this is a patient-centered and patient-engaged process.  

  • Probably the missing link here is the available data in structured, semantically interoperable form. Once available, SMART should be useable for data summarization. There appears to be sufficient momentum through federal funding sources such as NIH, AHRQ, CDC in this area, as well as other non-grant resources.  

  • Architecture and standards for semantically interoperable data are still not available, adopted, or verifiable to the extent needed.  

  • Terminology standards and resources are needed for relevant concepts. |
### Establishing a clear business case and providing incentives

<table>
<thead>
<tr>
<th>Management, but they have limited abilities to contribute data back in/write due to API limitations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The current system as it exists is generally a reflection of the current business incentives. In other words, the system we have now is optimized for the incentives we have now.</td>
</tr>
<tr>
<td>Cost of implementing desired solutions can be an important barrier. There needs to be a financial ROI case for the needed CDS and other HIT solutions, but such ROI information may not be clear or even available.</td>
</tr>
<tr>
<td>Change is happening in terms of overall incentives to incorporate CDS and improve quality:</td>
</tr>
<tr>
<td>- Value-based care should result in an incentive shift toward care that is better for opioid use and desired actions for opioid management.</td>
</tr>
<tr>
<td>- Pharma incentives to be seen as helping with this issue.</td>
</tr>
<tr>
<td>- Providers are motivated to avoid regulation-related, penalties, lawsuits, maintain licensure, etc.</td>
</tr>
<tr>
<td>- Many state medical boards are starting to put in opioid related requirements</td>
</tr>
<tr>
<td>Incentives are still limited/do not support improving pain management and opioid management. CMS is important, but private payors are also getting actively engaged in this area. We need quality measures around what is desired.</td>
</tr>
<tr>
<td>One potentially promising approach would be coming up with best practice recommendations that can be adopted by state medical boards and supported by technology.</td>
</tr>
<tr>
<td>Need payments for desired outcomes and potentially use of appropriate CDS tools (processes), including extra time required by clinicians for optimized patient care.</td>
</tr>
</tbody>
</table>

### Effective and Useful CDS

| What makes CDS effective is still not adequately defined and CDS interventions still often fail to achieve the desired outcomes. |
| The effectiveness of the vast majority of CDS interventions is never actually measured and it is unclear whether CDS interventions that are effective in one setting or clinical area would be useful in another setting or clinical area. Effectiveness of CDS may not be seen unless or until it exists at scale. |
| CDS platforms, whether standards-based or EHR vendor-based, may have many limitations to effectiveness. |
| There are a number of CDS best practice approaches published already. |
| There are also guides on IT usability that address issues such as access, uptake, adherence, and effectiveness, e.g., the ONC EHR usability change package. |
| Behavioral nudges have been shown to be important and may be fairly easy to scale for impact, e.g., the default number of opioid prescriptions dispensed. |
| There are a number of guidelines available to serve as the basis for CDS. |
| There is a need to continue to advance research to identify factors that contribute to CDS effectiveness and improve reporting of CDS intervention results. |
| Users of CDS platforms could help define desired/needed enhancements and join together in asking for those enhancements to be made. Incentives to develop more effective CDS are needed. |
| Sharing of CDS and good examples of effective CDS could allow for higher-quality CDS. |
| There is a lack of agreement on documentation templates/expectations in this space, an authoritative body should step forward to develop consensus on this. |

### Scaling CDS

| Scalable clinical decision support and knowledge sharing, whether via shared knowledge artifacts, or applications, or as services, is currently in its infancy. Common interoperability architecture also needs to be defined and adopted, e.g., SMART on FHIR, CDS Hooks. |
| There are local factors that need to be accounted for such as: business needs, practice settings, clinician types, patient populations, workflow approaches, documentation templates, and local resources. |
| There are a few computer-interpretable guideline repositories. |
| There is a white paper on how to establish trust in this ecosystem. |
| Some hybrid approaches exist where the “hard part” of guideline development is shared inter-institutionally, and each organization can adapt it to the local environment. |
| Standard value sets, such as those from the National Library of Medicine’s Value Set Authority Center. |
| Trust requirements are needed, especially for centrally managed CDS. |
| It may be useful to prioritize across domains to help ensure content quality e.g. the overall medication list, to consolidate guidelines whenever possible, and to better engage patients. |
Care Planning and Coordination

- There is a need for strong communication and shared mental models of care plans across organizations and providers, as well as between patients and their caregivers. Communication and information sharing among these providers can be suboptimal.
- Care plans are rarely synchronized while they evolve over time and are touched or updated by various stakeholders across organizational boundaries.
- Opioid use, a big enough, well recognized problem, may incentivize competing organizations to agree on common approaches to treatment and better coordination and cooperation.
- There is some ongoing HIE, TEFCA, etc. work on care planning and coordination, e.g., IHE models on dynamic care plans, with somewhat limited implementation.
- There is a need for demonstration projects funded by government, payors, to investigate how pay-for-value-incented providers can improve intra- and inter-institutional care planning.
- There is a need for more research around creating understandable and technically portable care plans for patients, and for more dynamic care planning.

Table 2. Derived insights for advancing PCCDS in general

<table>
<thead>
<tr>
<th>Category</th>
<th>Barriers</th>
<th>Facilitators</th>
<th>Needed Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulatory environment</td>
<td>It is challenging to harmonize across different regulatory domains and keep CDS information up to date.</td>
<td>Need accessible systems that link up regulatory resources and to maintain and update these centrally.</td>
<td>Governmental support of central management and use of source systems and data, including EBM and regulatory information, can buoy the data that underlies CDS.</td>
</tr>
<tr>
<td>Data integration</td>
<td>We rarely collect disparate source data with a fully integrated end in mind resulting in incomplete, missing, or incorrect data, limiting the effectiveness of CDS.</td>
<td>Various semantic interoperability standards and approaches can be used to improve the quality of source data and patients can assist with this.</td>
<td>Need a unifying interoperability specification to facilitate high-quality data collection informed by patient feedback.</td>
</tr>
<tr>
<td>Establishing a clear business case for CDS</td>
<td>The bulk of the challenge in establishing the ROI for CDS results from the current incentive system and the cost of implementing CDS.</td>
<td>Merit-based incentives and greater interoperability will help catalyze CDS integration.</td>
<td>Linking merit-based payments to quality measures, desired outcomes, the use of evidence-based practices, and/or the integration of CDS systems will establish a clear ROI for using CDS.</td>
</tr>
<tr>
<td>Effective and Useful CDS</td>
<td>CDS effectiveness remains poorly understood and implementations still suffer from usability (alert fatigue), adoption and information quality (trustworthiness) issues.</td>
<td>CDS best practices, usability best practices, quality guidelines, and behavioral nudges, are effective tools to support CDS implementations.</td>
<td>Engaging users in the specification, design, and implementation of CDS systems, creating publicly shareable success stories, and conducting research on factors that contribute to CDS efficacy and dissemination of CDS success stories can improve CDS implementations.</td>
</tr>
<tr>
<td>Scaling CDS</td>
<td>There is no common interoperability architecture to support the scalability of CDS.</td>
<td>Shareable repositories of CDS artifacts, guidelines, and standard value sets will improve CDS implementation at scale.</td>
<td>Emerging efforts to address trust for shared CDS will help support the expanded use of shareable CDS.</td>
</tr>
<tr>
<td>Care Planning and Coordination</td>
<td>The current environment does not support dynamic, multi-stakeholder, multi-environment care planning and coordination.</td>
<td>Need dynamic care plans and better engagement of the patient in the care planning process.</td>
<td>Need demonstration and research projects to focus on dynamic care planning (inter and intra-institutional) in the merit-based incentive context.</td>
</tr>
</tbody>
</table>
Effective and useful CDS: Making effective CDS both generally and specifically for opioids requires more insights. However, there is significant knowledge regarding CDS and usability best practice approaches. In addition, there is some evidence that simple behavioral nudges may also be effective. Needed Actions: The key is to identify and promulgate CDS successes and the factors that ensure success. To improve our insights into effective CDS, government should continue to fund research and researchers should continue looking for the most effective approaches to improving care using CDS. Increasingly, these efforts should engage users to provide the necessary input to ensure that CDS platforms meet user-centered design goals and vendors should help identify effective CDS and best practices and share these broadly.

Care planning and coordination: The complexity of patient care requires a shared understanding of the care planning process, including any integrated data available, among all members of the care team. The scale and impact of the opioid epidemic alone could help catalyze agreement on common approaches to coordination and treatment. Needed Actions: To mitigate this complexity, we need to develop an approach to cross-institutional care planning to better facilitate coordination. National-level funding for cross-institution demonstration projects and opportunities to incentivize value-based care will help. Researchers need to be encouraged to develop a body of evidence to guide the dynamic care planning process. We need relevant standards to help develop, pilot, and refine approaches to care planning and coordination.

Discussion

Summary of Findings. In 2018, the AHRQ PCCDS LN convened the TechFWG with the goal of identifying technical barriers to patient-centered CDS, with a particular focus on pain management. The TechFWG recruited 54 members, of which 21 members contributed substantially. Through 8 virtual meetings and 1 face-to-face meeting, the TechFWG members started with a vision for patient-centered CDS for pain management as defined by the PCCDS LN Opioid Action Plan TechFWG, then identified the barriers to realizing this vision at this time due to technical and related barriers. The core barriers were in the areas of 1) regulatory environment, 2) data integration, 3) scalability, 4) business case, 5) effective and useful CDS, and 6) care planning and coordination. Facilitators available for these barriers were identified, and recommendations were formulated for needed actions to address these barriers. When distilled into generalizable recommendations independent of the specific use case of pain management, the main recommendations from this work were as follows:

- Advocate to address conflict and overlap between federal, state, and local regulations.
- Define data needs and interoperability requirements with achievable goals tied more directly to clinical needs.
- Reach agreement on a vision of CDS at scale and develop specifications and implementations to reach this goal.
- Establish and promote the business case for CDS by incentivizing and educating stakeholder groups.
- Identify and broadly disseminate CDS success stories and related factors.
- Develop a standards-based approach to cross-institutional care planning to facilitate care coordination.

Strengths and Limitations. An important strength of this work was that by looking at the opioid use case in particular, the domain was specific enough to think through the implications at a good level of detail while the impact is broad enough (national-level) to facilitate generalization beyond...
this domain. A second strength is that we successfully convened a broad group of subject matter experts with insights and expertise in key domains, including standards-based interoperability, opioid and pain management CDS, and industry perspectives. Finally, a third strength is that the specific domain chosen is a critical one facing the nation, such that the insights and recommendations provided could have near-term impact on an important societal problem. With regard to limitations, one limitation is that the deliberations were based on a specific domain. Thus, it is possible that the barriers and associated recommendations are incomplete. However, many of the TechFWG members had broad experience in CDS in general, and the TechFWG explicitly sought to identify generalizable findings from their deliberations. Finally, as another related limitation, the focus on the pain management domain may have resulted in the inclusion of barriers that may not be as relevant in other patient-centered CDS use cases, such as regulatory barriers to data integration.

**Future Directions.** As the next step in this work, the PCCDS LN will be convening a Patient-facing CDS Demonstration Working Group (PFWG) which will seek to advance important components of the recommended actions developed in this work. The PFWG will tackle the patient-facing perspective of CDS and consider both design and technical factors that ease or inhibit the development of patient-centered CDS applications. The goals of the PFWG will be to actually convert an existing provider-facing SMART on FHIR CDS application based on the CDC guideline for morphine milligram equivalent dosing to a patient-facing application which can be directly accessed by patients, whether through the Web or a smartphone. In tandem with this technical development, the PFWG will also develop a set of human-centered usability design guidelines for patient-facing patient-centered CDS applications. This work will culminate in a set of recommendations for actionable steps that interested stakeholders could take to advance patient-centered CDS applications from prototype to use by patients, caregivers, and direct care clinicians.

CDS holds significant potential for improving the delivery of high-quality, patient-centered care, including for critical societal barriers such as the opioid epidemic. The findings from the PCCDS LN TechFWG could prioritize the needed actions for advancing patient-centered CDS to tackle these important barriers.

**Acknowledgments**

We acknowledge funding from AHRQ for its support via a Cooperative Agreement (U18 HS024849) of the PCCDS LN that provided the context and staff for developing this Technical Framework Report. In the past year, KK has been a consultant, sponsored researcher on clinical decision support, or invited speaker with honorarium for McKesson InterQual, Premier, Hitachi, the University of California at San Francisco, Klesis Healthcare, and the Office of the National Coordinator for Health IT. The other authors have no potential competing interests to declare.

**References**


Sharing of Individual Participant Data from Clinical Trials: General Comparison and HIV Use Case

Craig S. Mayer, MS1,2, Nick Williams, Ph.D1,2, Sigfried Gold, MFA, MA1, Kin Wah Fung, MD, MS1, Vojtech Huser MD, Ph.D1
1Lister Hill National Center for Biomedical Communication, National Library of Medicine, NIH Bethesda, MD
2ICF Incorporated, Rockville, MD

Abstract

Sharing of individual participant data is encouraged by the International Committee of Medical Journal Editors. We analyzed clinical trial registry data from ClinicalTrials.gov (CTG) and determined the proportion of trials sharing de-identified Individual Participant Data (IPD). We looked at 3,138 medical conditions (as Medical Subject Heading terms). Overall, 10.8% of trials with first registration date after December 1, 2015 answered ‘Yes’ to plan to share de-identified IPD data. This sharing rate ranges between 0% (biliary tract neoplasms) to 72.2% (meningitis, meningococcal) when analyzed by disease that is focus of a study. Via a predictive model, we found that studies that deposited basic summary results data to CTG results registry, large studies and phase 3 interventional studies are most likely to declare intent to share IPD data. As part of an HIV common data element analysis project, we further compared a body of HIV trials (24% sharing rate) to other diseases.

Introduction

Sharing of de-identified individual participant data1 is encouraged by the International Committee of Medical Journal Editors (ICMJE).2 Since 2019, authors that publish trial results in an ICMJE member journal (more precisely for trials that begin enrolling participants after Jan 1, 2019) are required to include a clear data sharing statement in the trial’s registration data2. This does not mandate actual individual participant data (IPD) sharing; it only mandates describing whether IPD data will or will not be shared. However, many funders of clinical studies have a separate policy that promotes actual data sharing.3 Data sharing is closely related to a trial registration mandate required for applicable trials in the US. ClinicalTrials.gov (CTG) is the world’s largest registry of interventional trials and observational studies4. CTG study record administrators are encouraged to keep the study record in the CTG registry current in order to facilitate public review of ongoing or completed studies. Since 2015, CTG allows study record administrators to specify an IPD data sharing plan.

We investigate the question of data sharing from the perspective of a study sponsor that is highly motivated to share data and consequently keeps the CTG study record current. Since 2017, CTG also added the ability to provide links to study protocol, Case Report Forms (CRFs), other study documents, or IPD. In case of documents, these can be uploaded directly to CTG or provided via a link. For IPD, only links are accepted. CTG’s ability to refer to other trial results platforms made CTG a central place and a good starting point to not only discover relevant clinical studies but to obtain important study documentation and IPD. We assumed that motivated study record sponsors would take advantage of CTG serving as the starting point to provide study metadata to the public. We fully acknowledge that the ability to provide a data sharing plan and provide an external link to de-identified IPD was only added in December of 2015 and trials that did first registration prior to this date were unable to do so at the time. However to assess the impact of the timing issue, we looked at how often such trials update the IPD data sharing plan retrospectively (during a registry record update). As of December 20, 2018, we found that a total of 18,457 studies (6.3% of 292,680 registered studies at that time) that initially registered when specifying a data sharing plan was not possible (prior December 1, 2015), that retrospectively answered ‘Yes’, ‘No’ or ‘Undecided’ to the question about an IPD data sharing plan. A total of 4,134 studies answered ‘Yes’, 10,379 answered ‘No’ and 3,944 answered ‘Undecided’. This finding shows that thousands of trials indeed update their CTG registration record, and it partially confirms its importance as a central public record.

In this study, we analyze study data sharing by disease domain. As part of a larger project focusing on analysis of clinical trials and their common data elements for the HIV and AIDS domain, we further focus on describing the body of all HIV trials (we use the term HIV trials to refer to both HIV and AIDS trials) and compare patterns of data sharing in HIV with other medical domains.
Methods

Determining data sharing on disease level

We used relational database representation of CTG study registration data called Aggregate Analysis of ClinicalTrials.gov (AACT) created at Duke University. The AACT database is created by parsing XML representation of each study from CTG. We adopt the CTG terminology and use the term study to refer to both an interventional trial and an observational study. We used SQL language to query and extract data from AACT and R language to analyze the data. The study repository at https://github.com/lhncbc/r-snippets-bmi/tree/master/CTG/sharing contains selected analytical code and supplementary result files.

To assign a given trial to a disease, we used CTG’s condition field. Each CTG study must specify a study condition that is defined as: ‘The name(s) of the disease(s) or condition(s) studied in the clinical study, or the focus of the clinical study.’ For example, the TOPKAT trial (NCT01352247) has ‘Knee Osteoarthritis’ as a single specified study condition. CTG aggregates similar condition entries into higher level groupings based on Medical Subject Headings (MeSH) terminology (CTG field: condition_mesh). The majority of trials (40.5%) have a single condition specified per study when all CTG studies are analyzed. A total of 29.9% of trials have two conditions and 29.6% have three or more conditions specified. Our intent was to analyze sharing by MeSH keyword, so studies that specified multiple keywords were included in any analysis for each specified keyword. To avoid bias in our analysis of overall sharing, we removed conditions that were considered outliers which had less than 15 registered studies.

To analyze data sharing intent, we used CTG’s field called plan_to_share_IPD. During study registration, the study record administrator must indicate whether there is a plan to make individual participant data (IPD) collected in the study, including data dictionaries, available to other researchers (typically after the study is completed and a possible data embargo period has expired). The data sharing plan has ‘Yes’, ‘No’ and ‘Undecided’ as possible answers. To avoid bias, we only looked at studies which first registered after December 1, 2015 when this feature was available during first registration, studies that provided an answer in the plan_to_share_IPD field, and studies that had an enrollment greater than zero (we refer to this set of studies as Set_A). For analysis purposes we viewed ‘Undecided’ as not stating that the study plans to share and so included it as a ‘No’ answer. A related second field in CTG is plan_to_share_IPD_description, which is a free text field where details of the plan are described. If the answer was ‘No’ or ‘Undecided’, the description field can be used to explain why IPD will not be shared or why it has not been decided yet.

Predicting Sharing

To explore what factors influence a decision to share study IPD, we developed a predictive model. The model used logistic regression analysis due to the categorical nature of the outcome variable. The predicted outcome variable was the answer of ‘Yes’ to the plan_to_share_IPD field in CTG. In order to select predictors, we first conducted a descriptive analysis of available CTG study level metadata for studies that answered ‘Yes’. Based on this analysis, we subsequently selected a set of predictors either used in past analyses of CTG data or identified in a review by two clinical informatics experts. The prediction model was done in an exploratory mode. A comprehensive model was out of scope.

HIV use case

Because of a larger project that focuses on common data elements in HIV, we analyzed as a case study the HIV body of trials. We did several HIV-specific analyses.

Sharing ratio: First, we created a set of MeSH terms that belong to the HIV/AIDS domain (reviewing only MeSH terms present in CTG) and compared HIV sharing ratios to other diseases.

Manual review of data sharing plan description: Second, we analyzed more closely the other data sharing CTG fields for HIV trials. CTG specifies the plan description should describe what specific individual participant data sets will be shared and how they will be shared. We only did this type of review for HIV studies (due to feasible sample size) and not for all CTG studies in Set_A. An example of a plan description is ‘de-identified participant level data will be available through the clinicalstudydatarequest.com platform upon request’. We manually reviewed plan descriptions (CTG field name: plan_to_share_IPD_description) for the 69 HIV studies that said ‘Yes’ to plan_to_share_IPD with a goal to assess how well the meaning of the IPD sharing fields were being understood by CTG study record administrators and how often a well specified plan is being articulated. Elements of a good plan were clearly specified by ICMJE. We also looked at instances and common answers where the fields were possibly misunderstood and
Predicting sharing (for HIV): We used the same predictive model approach to assess trial factors that predict IPD data sharing for HIV trials. We contrasted how HIV-specific model predictors are different from the general model (for all trials).

Results

We queried the AACT database on December 20, 2018 and analyzed 287,626 studies that were registered on CTG at that time. Of those registered studies, 74,582 provided an answer to the plan_to_share_IPD question. Overall, 9,435 studies (out of 287,626, 3.3%) are sharing IPD (without regard to first submission date or MeSH disease term for study focus), 46,017 studies (16.0%) are not sharing IPD and 19,130 studies (6.7%) were undecided on sharing IPD.

Determining data sharing on disease level

When we filtered the above set of studies with our selection criteria of (1) only include those registered after December 1, 2015 (the date when the CTG registry introduced additional data sharing questions), (2) that provided some answer (Yes/No/Undecided) to plan_to_share_IPD, and (3) had listed the enrollment of greater than zero, the number of analyzed studies reduced to 62,166 studies (21.6% out of 287,626; referred to as set_A). Supplemental file S1_diseases.xlsx (Tab1-all_keywords) at our study repository shows sharing ratio for all 3,138 MeSH keywords in this set of studies. The overall sharing ratio (without regard to MeSH term) for studies meeting these criteria is 10.8%. To better review sharing rate across diseases, we have removed outlier MeSH disease terms that have fewer than 15 total studies (e.g., adenocarcinoma in situ and coronary aneurysm; see Tab2-without-rare-MeSH-terms in the S1 file). The two MeSH terms with the highest sharing ratios were ‘meningitis, meningococcal’ (sharing rate of 72.2%) and dengue (61.1%). The lowest sharing ratio was zero, where no studies for those MeSH terms planned to share IPD. Zero sharing rate was observed for 70 MeSH terms. Example MeSH terms with a zero sharing rate and the highest total count of studies were biliary tract neoplasms (0 sharing out of 51 total studies), bone diseases (0 sharing out of 43 total studies), and obstetric labor, premature (0 sharing out of 39 total studies). A review of sharing rate and total count of studies for MeSH keywords with the largest count of total studies (after December 1, 2015), reveals the following MeSH keywords: diabetes mellitus (16.5% sharing out of 1,656 total studies), breast neoplasms (9.9% sharing out of 1,068 total studies) and depression (13.6% sharing out of 825 total studies).

Table 1. Share rate of a subset of diseases (by MeSH term)

<table>
<thead>
<tr>
<th>MeSH Term</th>
<th>Count of studies sharing data</th>
<th>Total count of eligible studies</th>
<th>Sharing ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>meningitis, meningococcal</td>
<td>13</td>
<td>18</td>
<td>72.2%</td>
</tr>
<tr>
<td>dengue</td>
<td>11</td>
<td>18</td>
<td>61.1%</td>
</tr>
<tr>
<td>hemophilia b</td>
<td>11</td>
<td>24</td>
<td>45.8%</td>
</tr>
<tr>
<td>carpal tunnel syndrome</td>
<td>21</td>
<td>64</td>
<td>32.8%</td>
</tr>
<tr>
<td>eczema</td>
<td>40</td>
<td>169</td>
<td>23.7%</td>
</tr>
<tr>
<td>diabetes mellitus</td>
<td>273</td>
<td>1656</td>
<td>16.5%</td>
</tr>
<tr>
<td>arthritis, rheumatoid</td>
<td>43</td>
<td>286</td>
<td>15.0%</td>
</tr>
<tr>
<td>depression</td>
<td>112</td>
<td>825</td>
<td>13.6%</td>
</tr>
<tr>
<td>asthma</td>
<td>33</td>
<td>327</td>
<td>10.1%</td>
</tr>
<tr>
<td>breast neoplasms</td>
<td>106</td>
<td>1068</td>
<td>9.9%</td>
</tr>
<tr>
<td>biliary tract neoplasms</td>
<td>0</td>
<td>51</td>
<td>0.0%</td>
</tr>
<tr>
<td>bone diseases</td>
<td>0</td>
<td>43</td>
<td>0.0%</td>
</tr>
<tr>
<td>obstetric labor, premature</td>
<td>0</td>
<td>39</td>
<td>0.0%</td>
</tr>
<tr>
<td>oropharyngeal neoplasms</td>
<td>0</td>
<td>35</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Predicting sharing

Descriptive analyses that were executed in support of selecting appropriate predictors are available in supplemental file S2-sharing-descriptive-results. The predictive model was executed for all studies in set_A with the predicted
outcome being an answer of ‘Yes’ to plan_to_share_IPD. The answer of ‘Undecided’ was essentially treated in the same group as an answer of ‘No’. Table 2 lists the relative effect of the predictors (their coefficients) and the predictor type (e.g., binary, categorical, continuous). Study sample size (CTG uses term enrollment to refer to study size in terms of number of enrolled participants) was converted into a categorical variable based on bin thresholds determined in the preparatory descriptive study (see section Study size in supplemental file S2). Large studies were considered having enrollment over 200, medium studies having enrollment of between 51 and 200 and small studies having enrollment of 50 or less.

To assess the predictors, the absolute value indicates its importance in the overall model. The positive sign of a predictor indicates higher probability of IPD sharing whereas negative sign indicates lower probability. Table 2 shows that the highest positive predictor is study size (studies with 201+ participants; coefficient 1.2) The second highest positive coefficient overall (has_results: Yes) was if the study has basic summary results deposited in CTG results registry. The highest negative predictors are sponsor types (categorized under Agency_class) of Other and U.S. Federal Agency. Outside of these sponsor types the two predictors with the largest negative coefficients, indicating a lower probability of sharing IPD were an observational study type and if the phase is listed as N/A, which includes studies on devices and behavioral interventions. While study type did not have as large an effect as other analyzed predictors on whether a study would share, observational studies and patient registries do have a negative effect on the probability of a study sharing.

In terms of terms of study sponsor, NIH sponsored studies are the most likely to share and has the highest positive effect of any sponsor type on the probability of a study sharing (0.296). The only other sponsor type with a positive effect on sharing is Industry (0.031). The other two sponsor types, U.S Federal agencies (non-NIH) and Other, had the greatest negative coefficients overall, indicating a significantly lower probability of IPD sharing by studies with these sponsors. Other, which includes foreign institutions, universities, hospitals and non-profit organizations, had the highest negative effect (-0.617) while U.S. Federal agencies, such as the VA and FDA had the second highest negative effect (-0.556). One surprise in the result or the coefficients of the predictor was that a study being phase 1 was not a negative predictor but rather had the highest positive effect of any phase on the probability of a study sharing.

We executed the prediction model for each MeSH term and the results are available in supplemental table S3-prediction-by-MeSH-term.

**Table 2.** Prediction model results for all studies showing predictor coefficients (ordered by absolute value)

<table>
<thead>
<tr>
<th>General Model Predictor</th>
<th>Predictor Coefficient</th>
<th>Predictor Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrollment: 201+</td>
<td>1.207541</td>
<td>categorical(3 values)</td>
</tr>
<tr>
<td>Has_results: Yes</td>
<td>0.977774</td>
<td>binary</td>
</tr>
<tr>
<td>Agency_class: Other</td>
<td>-0.617322</td>
<td>categorical(4 values)</td>
</tr>
<tr>
<td>Agency_class: U.S. Fed</td>
<td>-0.556474</td>
<td>categorical(4 values)</td>
</tr>
<tr>
<td>Phase: N/A</td>
<td>-0.396123</td>
<td>categorical(7 values)</td>
</tr>
<tr>
<td>Study_type Observational</td>
<td>-0.348703</td>
<td>categorical(3 values)</td>
</tr>
<tr>
<td>Enrollment: 51-200</td>
<td>0.325061</td>
<td>categorical(3 values)</td>
</tr>
<tr>
<td>Phase: 1/2</td>
<td>-0.322666</td>
<td>categorical(7 values)</td>
</tr>
<tr>
<td>Agency_class: NIH</td>
<td>0.296163</td>
<td>categorical(4 values)</td>
</tr>
<tr>
<td>Phase: 1</td>
<td>0.248599</td>
<td>categorical(7 values)</td>
</tr>
<tr>
<td>Phase: 2/3</td>
<td>-0.245158</td>
<td>categorical(7 values)</td>
</tr>
<tr>
<td>Phase: 3</td>
<td>0.227539</td>
<td>categorical(7 values)</td>
</tr>
<tr>
<td>Study_type: Observational [Patient Registry]</td>
<td>-0.142474</td>
<td>categorical(3 values)</td>
</tr>
<tr>
<td>Phase: 4</td>
<td>-0.125912</td>
<td>categorical(7 values)</td>
</tr>
<tr>
<td>Phase: 2</td>
<td>-0.104619</td>
<td>categorical(7 values)</td>
</tr>
<tr>
<td>Start_date</td>
<td>-0.071769</td>
<td>continuous</td>
</tr>
<tr>
<td>Agency_class: Industry</td>
<td>0.031220</td>
<td>categorical(4 values)</td>
</tr>
<tr>
<td>Enrollment: 1-50</td>
<td>-0.011347</td>
<td>categorical(3 values)</td>
</tr>
</tbody>
</table>

**HIV use case**

Sharing ratio: Table 3 shows all HIV/AIDS MeSH terms found in CTG and their sharing ratios. It shows that for the largest HIV MeSH term (HIV infections) the sharing rate was 24.5% which is more than double the average rate of
10.8% observed across diseases. The sharing rate for HIV places it in the top 5 percentile when looking at sharing ratio of all MeSH terms (when outliers are removed).

Table 3. Share rate for all HIV MeSH terms

<table>
<thead>
<tr>
<th>HIV MeSH Term</th>
<th>Count of studies sharing data</th>
<th>Total count of eligible studies</th>
<th>Sharing ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDS-related opportunistic infections</td>
<td>11</td>
<td>24</td>
<td>45.8</td>
</tr>
<tr>
<td>HIV seropositivity</td>
<td>7</td>
<td>25</td>
<td>28.0</td>
</tr>
<tr>
<td>HIV infections</td>
<td>69</td>
<td>284</td>
<td>24.3</td>
</tr>
<tr>
<td>AIDS dementia complex</td>
<td>1</td>
<td>6</td>
<td>16.7</td>
</tr>
<tr>
<td>HIV-associated lipodystrophy syndrome</td>
<td>0</td>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>lymphoma, aids-related</td>
<td>0</td>
<td>1</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Manual review of data sharing plan description: Since the inclusion of the plan to share IPD field in CTG there have been 69 studies with HIV related MeSH terms that have answered ‘Yes’ to plan_to_share_IPD for which we manually classified their data sharing plan descriptions. Of the 69 studies reviewed, 51 had well specified plans to share IPD, in which a specific platform for data sharing was stated or a process for which researchers can follow to obtain study data was articulated. There were 18 studies that had plans listed that were not well specified and were possibly misunderstood. Examples of common answers fitting this possibly misunderstood category were: (1) the field was left blank, (2) a plan will be developed in the future, (3) data will be shared via publications and/or presentations, and (4) data will be shared with the participants and investigators in the study. All of which are examples of plans that do not properly articulate a method to share IPD with interested parties and external researchers. In our 2017 prior work (across all diseases, not just HIV), we also noted that the plan was describing summary data sharing and not participant level data, which is another possible misunderstanding. The full list of the 18 misunderstood descriptions is available in the supplemental file S4-plans. CTG also allows trial administrators who answer ‘No’ or ‘Undecided’ to include an explanation of why they answered one of these two options, in the plan description field. Of the 215 HIV studies that answered ‘No’ or ‘Undecided’, 22 included an explanation in the plan_to_share_IPD_description field. Further showing possible misunderstanding of the IPD sharing fields, 7 of the 22 had descriptions of sharing plans that were well specified and included processes for outside researchers to obtain IPD. Such plan descriptions seem to imply ‘Yes’ to sharing; however, official CTG data had ‘No’ or ‘Undecided’ recorded.

Predicting sharing (for HIV): We executed the model using just the HIV set of studies as input and reviewed the results, focusing on the predicted effects of each variable for HIV studies compared to those of the overall model presented earlier.

Table 4. Prediction model results for HIV studies showing predictor coefficients (ordered by absolute value)
For HIV related studies, shown in Table 4, we found the most important predictors of sharing to be similar to those of the general model. The two highest positive coefficients were study size (studies with 201+ participants; coefficient of 2.01) and whether a study had basic summary results deposited in CTG. Compared with the general model, having summary results deposited (and several other predictors) had a much higher absolute value (‘has_results: Yes’ had 9.0 absolute value for HIV vs. 0.97 for the general model).

There were also some differences in the general versus HIV prediction model results. Certain factors that had a positive effect in the general model had a negative effect in the HIV model. This included the factors of both Phase 1 and Phase 3 as well as NIH sponsored studies. The largest negative factor on whether an HIV study would share IPD data was whether the study type was an observational patient registry. This is not a surprise as one major factor in sharing IPD is the completion of the study. Since patient registries do not have a limited temporal scope (are long term research projects) and are always actively adding data, the IPD dataset is technically never complete and has a different policy and sharing pattern than a typical interventional trial or time-limited observational study.

Another noticeable difference for the HIV model is the positive effect start date has on the sharing of HIV data compared to the minimal weight start date had on the general model, indicating a trend in increased sharing of HIV studies over time.

Discussion

Our study is the first to analyze intent to share IPD data by disease using CTG registration data. We show that there are significant differences in IPD sharing ratios across diseases (using MeSH terms). We also show different factors or predictors that impact IPD data sharing. The fact that studies that deposited basic summary results to CTG result registry are more likely to share IPD data is helpful in our larger project where we seek IPD data (or data dictionaries) for a large set of trials in a given disease (e.g., HIV, asthma or lymphoma). For trials where no answer to the question about IPD data sharing on CTG was provided (but basic summary results were deposited), a data outreach effort that first targets large trials with basic summary results may be the most pragmatic approach. Our study also shows that HIV studies share IPD at a higher rate (24.3%) than a general set of studies from the same time period (10.8%). HIV studies have a higher sharing rate despite the fact that the HIV-infected population may be considered a sensitive population.

Relevant previous studies

Prior to our analysis, we considered previous relevant literature on CTG data analysis and data sharing. Chen et al. analyzed digital health studies in CTG. Other analysis by Tse et al. and Durham look at challenges associated with CTG data and made recommendations on the use of CTG data. Stergiopoulos et al. reviewed the overall completeness of CTG records and the information present on CTG. Finally, Federer et al. analyzed data availability statements on a journal article level in a single journal PLOS ONE. To our knowledge, our study is the first to analyze (and compare across diseases) sharing of IPD data on study level using CTG registry data.

Limitations

The results presented have several limitations. First, in our analysis of sharing by disease, we used MeSH keywords and treated each MeSH term as a unique condition. We did not aggregate similar MeSH terms into possible higher level disease categories. This would require additional grouping knowledge base which we did not want to custom develop. This multiple keywords per disease can be observed, for example, on HIV keywords (see table 3) showing similar entries such as ‘hiv infections’ and ‘hiv seropositivity’. For HIV use case analysis, we also kept results on MeSH term level (without aggregation) and considered the term with the most studies to be representative of the domain. We also observed that CTG assigns some studies MeSH terms (as disease focus of the study) that are too broad (e.g., ‘syndrome’ or ‘disease’). We did not try to correct this misclassification. Second, when using the CTG field study_condition and assigning studies to a MeSH term, we did not try to select a single MeSH keyword for a study. We simply counted a study under each of its keyword. This double counting may be considered incorrect by some experts, however we chose to respect multiple keywords for a study as entered by trial record administrators. In other words, our chosen unit of analysis was MeSH keyword and not a study. Third, for our prediction model, we only used predictors available in CTG and picked predictors based on descriptive analysis, prior literature and expert opinion. A comprehensive prediction analysis could possibly incorporate a larger set of predictors.
Conclusion

We found that 10.8% of studies plan to share IPD data (as declared on CTG) and compared data sharing ratio by disease. The most important study features that predict future IPD data sharing are large sample size and deposition of basic summary results. As a case study, we analyzed more closely sharing rate and detailed plan description for HIV/AIDS studies.

Acknowledgement: This research was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

References


6. AACT database download [Internet]. Available from: https://aact.ctticleantrials.org/download


8. Medical Subject Headings [Internet]. Available from: https://www.nlm.nih.gov/mesh


Identification and Ranking of Biomedical Informatics Researcher Citation Statistics through a Google Scholar Scraper

Allison B. McCoy, PhD1, Dean F. Sittig, PhD2, Jimmy Lin, PhD3, Adam Wright, PhD4

1Vanderbilt University Medical Center, Nashville, TN
2The University of Texas School of Biomedical Informatics at Houston, Houston, TX
3University of Waterloo, Waterloo, ON, Canada
4Brigham and Women’s Hospital, Harvard University, Boston, MA, USA

Abstract

To overcome limitations of previously developed scientific productivity ranking services, we created the Biomedical Informatics Researchers ranking website (rank.informatics-review.com). The website is composed of four key components that work together to create the automatically updating ranking website: 1) list of biomedical informatics researchers, 2) Google Scholar scraper, 3) display page, and 4) updater. The interactive website has facilitated identification of leaders in each of the key citation statistics categories (i.e., number of citations, h-index, and i10-index), and it has allowed other groups, such as tenure and promotions committees, to more effectively and efficiently evaluate researchers and interpret the various citation statistics reported by candidates. Creation of the biomedical informatics researcher ranking website highlights the vast differences in scholarly productivity among members of the biomedical informatics research community. Future efforts are underway to add new functionality to the website and to expand the work to identify top papers in biomedical informatics.

Introduction

Citation statistics have been used to measure scholarly production of researchers since 1964, when Eugene Garfield created the Science Citation Index.1 Since then, additional measures have been developed in attempt to quantify research productivity and scientific impact independent of a researcher’s field of study and years at work, and without inflation from a small number of highly cited articles. Some services have attempted to rank research productivity, including ResearchGate,2 SciVal,3 and Highly Cited Researchers.4 However, these services have limitations, such as reliance on proprietary metrics, inclusion of only a limited number of highly ranked researchers, and requirements of organizational commitment to commercial sites.

Google created a new paradigm for citation analysis with Google Scholar,5 an online, freely available, automatically updating scientific information resource. By allowing researchers to create their own profile page,6 complete with multiple bibliometric calculations (e.g., total citations, h-index, i-10 index), Google provides a potential new method that allows comparison across researchers. Comparisons of Google Scholar with Web of Science and Scopus have found variations in citation statistics, where Google Scholar was frequently found to be considerably higher than Web of Science and Scopus due to Google Scholar’s wider inclusion of conference papers and gray literature.7,8 Despite these limitations, the format and availability of Google Scholar Profiles prove advantageous.

To facilitate comparison across researchers, we created a ranking website based on Google Scholar profiles for biomedical informatics researchers.9 This site was based on code previously developed by one author [JL] and used to create ranking sites for information retrieval,10 human-computer interaction,11 and top computer science researchers12 using a straightforward website scraping approach. These sites computed additional metrics to normalize Google Scholar’s bibliometric measures by dividing each of them by the number of years since the researcher received his or her first citation (i.e., citations/year, h-index/year, i-10 index/year).13

Design Considerations

Our goal in creating the Biomedical Informatics Researchers ranking website was to produce a freely available, automatically updating information resource based on Google Scholar citation profiles for all individuals interested in the field of biomedical informatics. Creating this resource required us to:

- Identify a set of biomedical informatics researchers with publicly-available Google Scholar profiles.
- Develop efficient methods to scrape the Google Scholar citation profiles of this list of individuals and extract key citation metrics.
- Implement a method to render the ranked list of researchers along with the means to re-order and search the list.
- Produce a method to allow new biomedical informatics researchers to add their name and google scholar profile location to the ranking site.
- Develop a method to automatically update the ranking site with the latest Google Scholar results on a periodic basis.

![Table showing the top 15 biomedical informatics researchers as of March 11, 2019](image)

**Figure 1:** Top 15 biomedical informatics researchers as of March 11, 2019
System Description

A snapshot of the ranking website as of March 11, 2019 is depicted in Figure 1. The website is composed of four key components that all work together to create the automatically updating ranking site: list of researchers, Google Scholar scraper, display page, and updater.

The first is the list of biomedical informatics researchers. The list file is in JSON (JavaScript Object Notation) format, with name / URL pairs represented as:

“Full Name [FACMI] [FAMIA] [FIAHSI] [Collen Year]”; “Google Scholar URL”

where:

- “FACMI” is an optional indicator that designates that they are member of the American College of Medical Informatics (ACMI).14
- “FAMIA” is an optional indicator that designates that they are a fellow of the American Medical Informatics Association (AMIA).15
- “FIAHSI” is an optional indicator that designates that they are a member of the International Academy of Health Sciences Informatics (IAHSI).16
- “Collen Year” is an optional indicator that designates that they are a recipient of the Morris F. Collen Award, and the year of their award.17

The FACMI, FAMIA, FIAHSI, and Collen Year components are appended to the name to be displayed on the ranked list of researchers. The list was created initially through an iterative process that began with manual searches for known biomedical informaticians using Google Scholar. After approximately 100 researchers were identified, we realized that we needed an automated method to develop a more comprehensive list. Therefore, in 2014, we used the “label:biomedical_informatics” search feature that identified all individuals on Google Scholar with “biomedical informatics” as one of their “areas of interest” and at least one publication with one or more citations. We repeated this search using “label:medical_informatics” and other common key words related to biomedical informatics, including health informatics, electronic health record, clinical decision support, and health information technology.

To facilitate new requests to be added to the list of biomedical informatics researchers, we created a Google form that allows a researcher to requested to add his or her profile to the ranking website. The Google form prompts researchers to enter his or her name and Google Scholar URL; to indicate whether they are an ACMI fellow, AMIA fellow, or member of IAHSI; and to indicate whether they have received the Morris F. Collen award and, if so, to enter the year. The input data is manually verified to be accurate by one of the co-authors to prevent errors in running the scraper, then it is added to the list of biomedical informatics researchers to be included on the site with the next update. Since then, we have periodically solicited requests for individuals to add their profiles (or create one if they had not already done so) through the ACMI listserv and other targeted e-mailings (e.g., to department listservs, through AMIA Connect). We have also manually added new profiles found through repeated Google Scholar searches of relevant labels and review of new individuals listed on ACMI, FAMIA, and IAHSI member lists.

The second component is the Google Scholar Scraper. This open-source application is written in node.js and built using commonly-available open-source libraries. It takes as input the list of researchers and then iteratively retrieves the listing of each person’s google scholar citation counts, the total number of citations, the year of first citation, the i-10 index, and the h-index. These values are extracted based on matching the relevant elements from each page’s DOM (Document Object Model) structure. This approach makes the scraper application dependent on the layout of the Google Scholar profile page, so it is not robust to changes in the layout of the profiles, and indeed, the application has broken several times since the initial development in 2014 after Google updated its site. However, no APIs (Application Programming Interfaces) that allow programmatic access to such data are available, so there are few alternatives to this screen-scraping approach.

In addition to extracting raw statistics from profile pages, the application also calculates the citations/year, i-10 index/year, and h-index/year; all computed values are written into a file in JSON format, which facilitates the display as well as downstream processing by other applications. The following is a brief definition of each of the bibliometric measures included on the ranking site:

- Total number of citations – the total number of citations to all of a researcher’s published articles
• Year of first citation – the year in which the researcher received his or her first citation, regardless of the year of publication of their first article

• i-10-index – the number of articles that a researcher has published that have received at least 10 citations

• h-index – the number of articles (n) that a researcher has published that have received at least “h” citations where n=h. In other words, if a research has published 25 articles that have all received at least 25 citations, then his or her h-index is 25.

• Citations/year – a researcher’s total number of citations divided by the number of years in which he or she has been accumulating citations (i.e., current year – year of first citation)

• i10-index/year – a researcher’s i-10-index divided by the number of years in which he or she has been accumulating citations

• h-index/year – a researcher’s h-index divided by the number of years in which he or she has been accumulating citations

The third component is the display page that renders the JSON data created by the scraper program above in HTML/CSS, (Hypertext Markup Language / Cascading Style Sheets) with the aid of JavaScript. The display lists the researchers in ranked order and allows a user to re-sort the entire list by any of the column headers (e.g., citations or i-10 index). The display page also incorporates a search feature that allows one to display a ranked subset of researchers, for example: 235 ACMI members, the 32 people associated with “Vanderbilt University”, or the 26 people with “David” in either their name or affiliation. This page also includes code that allows Google Analytics to track website traffic.

The final component is the updater, a script that periodically re-runs the scraper and pushes the updated data to the web site and Github. The updater is currently set to run twice per week. Several error conditions are periodically encountered and detected by the updater script, including instances where people delete their Google Scholar profile or make it private, network issues that prevent connection to Google Scholar or Github, or temporary blocks imposed by Google. Although Google permits scraping of Google Scholar profiles in their robot exclusion standard (robots.txt) file, they do periodically block the scraper if it is set to run too often.

**Current Status of Biomedical Informatics Researcher Ranking Website**

The Citation Statistics of Biomedical Informatics Researchers ranking website can be viewed at rank.informatics-review.com. In the nearly five years since its inception, the website has been viewed more than 18,000 times by almost 9,000 users. Of these users, 70% reside in the United States, 6% in India, 2.5% in Australia, 2.4% in Canada, 1.7% in the United Kingdom, and 1% in China. We observed apparent spikes in website traffic in several instances after listserv e-mails were sent or individual researchers mentioned the website on social media (Figure 2). For example, timepoint #1 corresponds with an e-mail sent to the ACMI listserv, and timepoint #4 corresponds with a Tweet by @allisonbmc-coy.

![Figure 2: Visitors to the biomedical informatics ranking website from Google Analytics with traffic spikes corresponding to known instances of dissemination.](image)

The list of biomedical informatics researchers contains 1,401 individuals, including 235 ACMI fellows, 62 AMIA fellows, 61 IAHSI members, and 12 Morris F. Collen award winners. Requests to be added to the site have been submitted through the Google form for 171 researchers.
Since the BMI ranking list has been available, numerous uses for the list have been identified, including:

- To create a list of members from a single university and compare the scholarly productivity of those university’s biomedical informatics departments. To our knowledge, at least three universities are currently using the website in annual department reviews.19
- To identify productive researchers for nomination to ACMI or IAHSI members.
- To identify potential recruits for academic positions.
- To help tenure and promotions committees to interpret the various citation statistics reported by candidates.
- To identify speakers for conferences.
- To identify subfields of biomedical informatics for which citations are highest.20

**Notable Statistics for Biomedical Informatics Researchers**

Table 1 shows the median, min, and max values for all biomedical informatics researchers as well as for all ACMI, FAMIA, and IAHSI members and Morris F. Collen award winners, as identified through the biomedical informatics ranking website as of March 4, 2019. As expected, the median ACMI and IAHSI members (h-index=35.5 and 41, respectively) have been publishing for 8-10 years longer than the median for all researchers (h-index: 15). The median h-index for AMIA fellows (14) is similar to the median for all researchers, which is also expected given that FAMIA recognition is based on application of informatics skills and knowledge, regardless of research productivity. Table 2 shows the values for 10 randomly chosen Nobel Prize winning scientists (median h-index: 120) as an upper extreme for comparison.22

**Table 1: Descriptive analysis of citation statistics for biomedical informatics researchers**

<table>
<thead>
<tr>
<th>Year of 1st Citation</th>
<th>All Biomedical Informatics Researchers (N=1,401)</th>
<th>ACMI Fellows (N=235)</th>
<th>AMIA Fellows (N=62)</th>
<th>IAHSI Members (N=61)</th>
<th>Morris F. Collen Award Winners (N=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Citations</td>
<td>Median</td>
<td>1,028</td>
<td>5389</td>
<td>812.5</td>
<td>7145</td>
</tr>
<tr>
<td>Min</td>
<td>2</td>
<td>300</td>
<td>72</td>
<td>247</td>
<td>4,456</td>
</tr>
<tr>
<td>Max</td>
<td>166,410</td>
<td>166,410</td>
<td>24,324</td>
<td>108,929</td>
<td>108,929</td>
</tr>
<tr>
<td>Citations/ year</td>
<td>Median</td>
<td>68</td>
<td>237</td>
<td>55</td>
<td>274</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>10</td>
<td>5</td>
<td>15</td>
<td>117</td>
</tr>
<tr>
<td>Max</td>
<td>9,958</td>
<td>6,400</td>
<td>950</td>
<td>4,951</td>
<td>4,951</td>
</tr>
<tr>
<td>h-index</td>
<td>Median</td>
<td>15</td>
<td>36</td>
<td>14</td>
<td>41</td>
</tr>
<tr>
<td>Min</td>
<td>1</td>
<td>9</td>
<td>3</td>
<td>7</td>
<td>31</td>
</tr>
<tr>
<td>Max</td>
<td>199</td>
<td>149</td>
<td>72</td>
<td>149</td>
<td>149</td>
</tr>
<tr>
<td>h-index/ year</td>
<td>Median</td>
<td>1</td>
<td>1.6</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>Min</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Max</td>
<td>15.6</td>
<td>7.7</td>
<td>3.3</td>
<td>6.8</td>
<td>6.8</td>
</tr>
<tr>
<td>i10-index</td>
<td>Median</td>
<td>20</td>
<td>77</td>
<td>17.5</td>
<td>117</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>77</td>
</tr>
<tr>
<td>Max</td>
<td>922</td>
<td>802</td>
<td>252</td>
<td>695</td>
<td>695</td>
</tr>
<tr>
<td>i10-index/ year</td>
<td>Median</td>
<td>1.4</td>
<td>3.4</td>
<td>1.2</td>
<td>4.6</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Max</td>
<td>51.3</td>
<td>33.9</td>
<td>9</td>
<td>31.6</td>
<td>31.6</td>
</tr>
</tbody>
</table>

ACMI = American College of Medical Informatics, AMIA = American Medical Informatics Association, IAHSI = International Academy of Health Sciences Informatics
In reviewing the citation ranking statistics on the website (Figure 1) and changing the primary sort order by column, we have made a number of interesting observations about the list:

- Eugene Koonin from National Center for Biotechnology Information has the most citations (166,410) and the highest h-index (199).

- Twenty-nine researchers are tied for the earliest “year of first citation” (1980). In reviewing these results, we noted that this is a limitation in the Google Scholar profile page, with no citations depicted prior to this date, though prior publication dates are listed on individual researchers’ pages (e.g., Homer Warner, 195121).

- Alex Wang from National Institutes of Health has the highest i10-index (922).

- Brian Pollack from University of Pittsburgh has the highest citations/year (9,958), h-index/year (15.6), and i10-index/year (51.3).

- ACMI members make up 55 of the top 100 researchers when sorted by both h-index and citations, and IAHSI members make up 20/100.

![Figure 3: Graph showing the relationship between h-index and total citations ($r^2=0.77$).](image-url)
To evaluate the included citation statistics, we calculated the correlation coefficient between the h-index and total citations ($r^2=0.77$) (Figure 3) and i10-index ($r^2=0.89$) (Figure 4) using Stata/IC 15.1. Overall, the statistics similarly portray researcher productivity; however, in one case a researcher has a disproportionately high total citation count compared to h-index due to a single paper with more than 100,000 citations.

Lessons Learned

Creation of the biomedical informatics researcher ranking website highlights the vast differences in scholarly productivity among members of the biomedical informatics research community. Careful inspection of the citations included on many researchers’ profile pages also highlights many of the limitations of automatically curated lists, including:

- For individuals with relatively common family names, the inclusion of articles that were authored by other researchers are often included erroneously, which can falsely inflate citation statistics and rankings. Authors can curate their own profiles to remove erroneous citations, but few do.

- Duplicate citations exist in many profiles that can also false inflate citation statistics and rankings; however, Google Scholar has implemented functionality to automatically merge some articles and combine citations when authors do not manually merge citations.

- Researchers with publications before the 1990’s, when use of the internet substantially increased, are not as well included in the various citation statistics. Most notable is that there are no citations included before 1980 in any of the counts, an important limitation of Google Scholar profile page and the scraper tool.

- Highly cited publications by large consortia, including the “Initial Sequencing and analysis of the human genome” and “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC” heavily skew some authors’ citation statistics.

- Not all articles included are equal, although Google Scholar often lists blog posts and slide presentations with articles from peer-reviewed scientific journals.

- Likewise, not all citations are equal, although Google Scholar counts all citations equally, whether from a website, slide presentation, or top scientific journal.

- Most indexed articles are in English, which negatively affects non-English speaking researchers.
Future Directions

While we believe the current biomedical informatics researcher ranking site is already very useful, we are continuing to identify new researchers, especially those who are highly cited, ACMI fellows, members of IAHSI, or Morris F. Collen award winners. In addition, we are reviewing profiles with a large number of incorrect or duplicate citations and requesting that the individuals curate their profile or be removed from the list. We have also identified numerous enhancements that we hope to make in the future, including:

- Adding the total number of articles included in each person’s Google Scholar profile and the year of first publication to the biomedical informatics researcher ranking website.
- Adding an indicator for other noteworthy accomplishments, including AMIA signature awards (e.g., Donald A.B. Lindberg Award for Innovation in Informatics, Virginia K. Saba Informatics Award, and AMIA New Investigator Award).
- Calculating the longest consecutive string of years in which each researcher published one or more articles that received one or more citations.26
- Calculating $h$ (universal h-index), or the h-index of an individual divided by the mean h-index of everyone in the field.27
- Evaluating and improving the usability and efficiency of the site.

Finally, we are exploring opportunities to use the current Google Scholar scraper to identify top papers in biomedical informatics for all time and in the past year. A preliminary version of this new tool retrieved the top 100 most cited publications with 100 or more citations from all profiles in the list of biomedical informatics researchers and found 7,429 papers that met these criteria. The top most cited publication had 69,812 citations total with 3,173 citations per year.28 A preliminary version of the tool to identify top papers in the last year retrieved all publications in 2018 from all profiles in the list of biomedical researchers and found 3,751 publications. The top most cited publication had 2,177 citations total.29 At present, several limitations to this new tool exist. One important limitation is the inclusion of all publication types; in 2018, the most cited publication was a textbook. Another limitation is the inclusion of papers published by biomedical informatics researchers in areas that are not directly related to biomedical informatics; for example, an American Heart Association report on which a biomedical informatics researcher played a small role related to informatics development or data analysis is the second most cited publication in 2018.30

Conclusion

We have developed an easily searchable, interactive, automatically updating, open-source bibliometric ranking website using Google Scholar citation profiles that includes over 1,300 biomedical informatics researchers from around the world. While there are limitations to both using bibliometric citation analysis to measure scientific productivity and automatically generated lists of articles and citations, the biomedical informatics ranking website has already proven to be useful for a number of important tasks. Future efforts are underway to add new functionality to the website and to expand the work to identify top papers in biomedical informatics.

References

FHIR OWL: Transforming OWL ontologies into FHIR terminology resources

Alejandro Metke-Jimenez, PhD, Michael Lawley, PhD, David Hansen, PhD
The Australian eHealth Research Centre, CSIRO, Herston, Queensland, Australia

Abstract
The FHIR specification provides a mechanism to access clinical terminologies using a standard API, and many existing terminologies, such as SNOMED CT, are well supported. However, in areas such as genomics, terminologies from other domains are starting to be used in clinical settings. Many of these are authored or distributed in Web Ontology Language (OWL) format. In this paper we describe a transformation between OWL ontologies and FHIR terminology resources. The results show that there are several challenges in implementing the transformation, with the major one being the lack of a modularisation mechanism in the FHIR code system resource that resembles the import mechanism available in OWL. A workaround with minimal drawbacks was successfully implemented in this solution. The availability of this transformation is significant because it enables a broad range of terminologies that are currently available in OWL to be available using the FHIR API.

Introduction
The Fast Healthcare Interoperability Resources (FHIR) standard defines a set of terminology resources that provide a mechanism to access clinical terminology in a standardised way. Many clinical terminologies are currently supported, including SNOMED CT, RxNorm and LOINC. Recently, there has been increasing interest in supporting terminologies from other domains. In genomics, for example, there is increasing use of the Human Phenotype Ontology (HPO) to capture information about rare diseases.

Many terminologies in other domains are authored or distributed using the Web Ontology Language (OWL), including HPO and NCI Thesaurus. This paper describes how to transform OWL ontologies into FHIR terminology resources and uses HPO as a case study. The challenges in the transformation are explained and a detailed overview of the mapping between OWL and FHIR is provided. The resulting transformation is generic and can be used to transform other OWL ontologies.

Having HPO available as a FHIR code system allows it to be accessed through the same FHIR-based API as SNOMED CT and other clinical terminologies, enabling OWL ontologies such as HPO to be easily supported by clinical information systems. HPO together with SNOMED CT can provide a more comprehensive set of concepts to describe clinical findings.

1 Background and Related Work
To the best of our knowledge, no other transformation between OWL ontologies and FHIR terminology resources has been published in the literature, with the exception of a paper on Ontoserver, a FHIR terminology server, that describes a preliminary version of this transformation.

In the following sections we review some basic concepts around FHIR terminology resources and OWL ontologies.

1.1 FHIR Terminology Support
The FHIR resources required to support terminology in FHIR are CodeSystem, ValueSet, ConceptMap, NamingSystem and TerminologyCapabilities. A detailed description of the FHIR terminology subsystem can be found in the FHIR specification. This work is based on FHIR R4, which is the current version at the time of writing. The transformation presented in this paper targets only the CodeSystem resource.

In FHIR, a code system represents a set of codes from a system. Each terminology is represented by a code system resource within a FHIR server. Some examples of large, standardised terminologies are SNOMED CT, RxNorm and LOINC. Note that even though the authoring and distribution formats for these terminologies are very different, its content can be accessed in a uniform way using the FHIR API. Users can also create and upload custom code systems
that are prevalent in the healthcare domain. More details about the CodeSystem resource can be found in the FHIR documentation available at http://hl7.org/fhir/codesystem.html.

1.2 OWL Ontologies

OWL is a language designed to represent complex knowledge about things. This includes properties of things, groups of things and relations between things. This knowledge, called an ontology, is meant to evolve constantly. The things described in an ontology are referred to as individuals and the relations as property assertions, which relate these individuals to each other. The groups of things are referred to as classes. An ontology is made of axioms that constrain how the individuals relate to each other through the definition of classes. This set of axioms can be interpreted by a software component, called a reasoner, that automatically infers new knowledge. An interesting example in the medical domain can be found in the work by Baadar, Lutz and Suntisrivaraporn.

There are three sublanguages in OWL 2 that support different levels of expressivity: OWL EL, OWL QL and OWL RL. These sublanguages limit the constructs that can be used when authoring an ontology, which reduces the kind of knowledge that can be expressed but facilitates the implementation of efficient reasoning algorithms. Most ontologies in the biomedical domain fall into the OWL EL profile.

2 Methods

To determine how to transform OWL ontologies into FHIR CodeSystem resources, an in-depth analysis of the key similarities and differences between both representations was conducted. Feedback about the different alternatives was also sought from FHIR experts through the FHIR community chat.

The FHIR resource that aligns more closely with OWL ontologies is CodeSystem, which is used to define codes that are relevant in the healthcare domain. This resource provides a simple mechanism to access codes in a standardised manner, deal with different versions, specify properties for each code and define filters based on these properties. However, FHIR code systems are not intended to support the formal definition and maintenance of large code systems, but rather to support standardised access to these and the creation of small, ad-hoc code systems. Many code systems supported in FHIR, such as SNOMED CT and LOINC, for example, have their own distribution formats and maintenance processes.

OWL, on the other hand, is much more complex than the FHIR terminology resources and is typically used to author ontologies with the aid of authoring environments such as Protege.

Most of the properties available in a FHIR code system can be found in an OWL ontology. Table 1 shows the properties of a code system and which ones have default values in OWL. These default values can always be overridden to use either a value passed in by the user or to use a different annotation property to look for values in the ontology.

<table>
<thead>
<tr>
<th>Code System</th>
<th>Default Values</th>
<th>Option</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>-id</td>
<td></td>
<td>The technical id of the code system. Required if using PUT to upload the resource to a FHIR server.</td>
</tr>
<tr>
<td>language</td>
<td>-language</td>
<td></td>
<td>The language of the content. This is a code supplied by the user from the FHIR Common Languages value set (<a href="https://www.hl7.org/fhir/valueset-languages.html">https://www.hl7.org/fhir/valueset-languages.html</a>).</td>
</tr>
<tr>
<td>url</td>
<td>Ontology IRI</td>
<td>-url</td>
<td>Canonical identifier of the code system. If this option is not specified then the ontology’s IRI will be used. If the ontology has no IRI then the transformation fails.</td>
</tr>
<tr>
<td>identifier</td>
<td>-identifier</td>
<td></td>
<td>Additional business identifiers.</td>
</tr>
<tr>
<td>Property</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>version</td>
<td>Business version. If this option is not specified then the ontology’s version will be used. If the ontology has no version then the version is set to ‘NA’.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>Computer-friendly name. If this option is not specified then the value will be taken from the OWL file using the RDFS:label property by default. The property to use can be overridden using the name-Prop option. Otherwise the ontology’s IRI is used.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Human-friendly name.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>status</td>
<td>Code system status. Valid values are: draft, active, retired and unknown.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>experimental</td>
<td>Indicates if the code system is for testing purposes or real usage.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>date</td>
<td>The published date.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>publisher</td>
<td>The publisher of the code system. If a publisher is not supplied using the publisher option, then the value will be taken from the annotation property in the OWL file specified using the publisherProp option. If this option is not present then no publisher is assigned.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>contact</td>
<td>Contact details for the publisher.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>description</td>
<td>The description of the code system. If a description is not supplied using the description option, then the value will be taken from the annotation property in the OWL file specified using the descriptionProp option. If this option is not present then no description is assigned.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>useContext</td>
<td>Not supported.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jurisdiction</td>
<td>Not supported.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>purpose</td>
<td>Explanation of why this code system is needed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>copyright</td>
<td>A copyright statement about the code system.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>caseSensitive</td>
<td>Always set to true because OWL is case sensitive.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>valueSet</td>
<td>The value set that represents the entire code system.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hierarchyMeaning</td>
<td>Always set to is-a because that is the hierarchy meaning in OWL.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>compositional</td>
<td>Flag to indicate if the code system defines a post-coordination grammar.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>versionNeeded</td>
<td>Flag to indicate if the code system commits to concept permanence across versions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>content</td>
<td>The extent of the content in this resource. Defaults to complete but can be set to other values. The actual value does not affect the output of the transformation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>supplements</td>
<td>Not supported.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>count</td>
<td>Calculated and assigned automatically.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Three filters are added by default: root, deprecated and imported.

Four properties are added by default: parent, imported, root and deprecated.

Table 1: Code system elements and their corresponding elements in OWL ontologies.

A concept in a FHIR code system typically corresponds to a class in OWL, which defines a “group of things”. Table 2 shows how the elements of an OWL class are mapped into a FHIR concept. The IRI of the class is mapped to the concept’s code by default, but this can be overridden so an arbitrary annotation property can be used instead. In HPO, the default mapping works well but in other cases the code is contained in other properties. For example, in the OWL version of the HUGO Gene Nomenclature, the code does not correspond to the IRI but rather to the HGNC_ID annotation property.

Table 2: Code system concepts and their corresponding elements in OWL ontologies.

2.1 Challenges

The are three main challenges when designing the transformation. The first is the absence of an import mechanism in FHIR code systems that is equivalent to the import mechanism available in OWL ontologies. When authoring an OWL ontology, external ontologies can be imported and existing concepts can be used to model new concepts, promoting modularity and reuse. For example, the Human Phenotype Ontology (HPO) imports 11 additional ontologies, some of which also import other ontologies. In FHIR, code systems do not support importing external code systems. The only mechanism that has been designed to include content from external code systems are code system supplements, which allow defining code systems that add designations and properties to existing code systems. However, new concepts cannot be defined in the supplementing code systems. Overall, there is no in-built mechanism to import external code systems in a way that is similar to the importing mechanism in OWL. Interestingly, the CTS2 specification, which predates FHIR and informed some of its terminology features, supports an includes attribute that does correspond to an OWL import statement.

The second challenge is determining which concepts belong to the main ontology and which ones are imported. This appears to be a simple problem but it is not straightforward for several reasons. First, an OWL class that was defined in an imported ontology can be redeclared in the main ontology with no modifications. In this case, it is clear that the class belongs in the imported ontology. However, it is also valid to do certain modifications to imported classes. In this case, the class could be considered part of the main ontology. Also, it is valid to declare classes in an ontology

---

1The IRI does resemble the id but it is formatted slightly differently, with an underscore instead of a colon, e.g., HGNC_100 instead of HGNC:100.
document with multiple namespaces, so the base namespace is not enough to determine if a class is imported or not. In certain cases, for example in HPO and other ontologies developed by the OBO Foundry, a naming convention is used, which makes it easy to determine if a class is imported or not, but not all ontologies use similar conventions so it is not feasible to rely on these in a generic solution.

Finally, FHIR code systems are meant to both support standardised access to existing code systems and the authoring of small, ad-hoc ones. The standard is not meant to support the authoring and distribution of existing code systems that use their own authoring and distribution mechanisms. For example, SNOMED CT uses RF2 to distribute its content and the FHIR standard does not describe how to process it, only the way users should access it using the FHIR terminology resources. Implementers are free to design their own mechanisms to process RF2 files and import the code system into their server. Transforming OWL ontologies into FHIR code systems breaks this assumption, because OWL ontologies can be very large. Existing terminology servers designed to deal only with small FHIR code systems might require modifications to perform well with FHIR code systems of this size.

2.2 Proposed Solutions

To overcome the first challenge, the mismatch between the import mechanism available in OWL and similar mechanisms in FHIR code systems, three alternatives were evaluated. First, each OWL ontology was created as a separate code system and the relationships between concepts that spanned ontologies were defined using a property, external-parent, of type coding. The coding type in FHIR allows referring to a code in a specific code system, while the code type does not support specifying a system and therefore assumes that the code belongs to the current code system. The main issue with this approach is that the parent and child properties available by default are of type code and therefore are not designed to reference external code systems. The main implication of using a non-standard property is that none of the subsumption features in FHIR, such as search by subsumption, will work. An additional drawback is that sometimes the ontologies that get imported into the main ontology are not complete, but rather just import modules, designed to contain only a subset of the ontology. This is typically done for performance reasons if the imported ontology is very big. Some of these import modules refer to their source ontology through the Dublin Core source property (for example, the Phenotypic Quality Ontology import module used in HPO) but some others do not (for example, the Relation Ontology import module, used recursively in HPO). When creating code systems for these, it is important to know if these ontologies represent fragments or complete code systems in order to set the content property correctly. The absence of a reference to the full ontology in import modules will result in a generated code system that claims to be complete.

The second approach creates a single code system with all the concepts in the ontology closure (i.e., all the concepts in the base ontology plus all the concepts in the imported ontologies, including recursive imports). A value set is created for each ontology so searching can be constrained to any single ontology. This approach overcomes the issue of having to use a bespoke property to build the hierarchy. Now the standard parent and child properties can be used because all the codes belong to the same code system. However, the main drawback of this approach is that the concepts’ codes need to include the namespace (i.e., they need to be the long form of the IRI) because otherwise there is the possibility of having name clashes. Also, strictly speaking, the definitions are incorrect because external codes are specified as belonging to the main system, when in reality this is not the case.

The final approach also creates a single code system but treats the concepts from the main ontology differently from the imported concepts. The property imported is created to flag if a concept belongs to the main code system or is imported from an external code system. Also, the imported concepts include the namespace in their codes, similarly to the second approach, but the concepts that belong to the main code system only use the short form of the IRI because the namespace is the system of the main code system. This approach still allows using the standard code system properties parent and child to build the hierarchy but has the advantage of clearly indicating that some concepts used in the code system are defined externally. Both the value of the imported property and the value of the code can be used to determine if the concept is imported. The main limitation of this approach is that it is possible to create redundant instances of imported concepts if multiple ontologies that import the same ontologies are used. So, for example, if both HPO and my new ontology use the Foundational Model of Anatomy (FMA), then all the concepts from FMA used in both ontologies will be created, which results in a waste of space. However, we believe the benefits
of this approach outweigh this relatively minor issue and therefore this was the solution adopted in the transformation. Figure 1 shows an example of a local and and imported concept when using this approach. The local concept, Abnormal cardiac septum morphology from HPO, uses the short form of the code and has the imported attribute set to false. The imported concept, septum from UBERON which is used in HPO, uses the long form of the code and has the imported attribute set to true.

```json
"concept": [  
  {  
    "code": "HP:0001671",  
    "display": "Abnormal cardiac septum morphology",  
    "property": [  
      {  
        "code": "imported",  
        "valueBoolean": false  
      },  
      {  
        "code": "parent",  
        "valueCode": "HP:0001627"  
      },  
      {  
        "code": "root",  
        "valueBoolean": false  
      },  
      {  
        "code": "deprecated",  
        "valueBoolean": false  
      }  
    ]  
  }]

"concept": [  
  {  
    "code": "http://purl.obolibrary.org/obo/UBERON_0003037",  
    "display": "septum",  
    "property": [  
      {  
        "code": "imported",  
        "valueBoolean": true  
      },  
      {  
        "code": "parent",  
        "valueCode": "http://purl.obolibrary.org/obo/UBERON_0000061"  
      },  
      {  
        "code": "root",  
        "valueBoolean": false  
      },  
      {  
        "code": "deprecated",  
        "valueBoolean": false  
      }  
    ]  
  }]
```

Figure 1: The FHIR representation of a local (left) and and imported (right) concept using the proposed approach.

This provides a good segue into the solution for the second challenge, determining which concepts belong in the main ontology and which ones are imported. The most straightforward alternative is to accept a user-defined list of namespaces that indicate if a concept belongs in the main ontology. Any concepts that use other namespaces will be considered imported. This strategy is implemented relaxing the definition of namespace by treating them as prefixes. In HPO, providing a list with the single element, http://purl.obolibrary.org/obo/hp_, produces correct results. If no input is provided then another alternative is to consider the classes defined in the main ontology document as local and all the others as imported. This approach does not work properly with HPO because several classes are redeclared in the main document, for example, several classes from the Chemical Entities of Biological Interest (ChEBI) ontology. Finally, a more sophisticated approach that does not require user input is to classify as local all the classes declared in the main document minus the ones declared in the imported documents. For HPO, this approach produces a result that is very close to being correct. The only issue is that some concepts are incorrectly classified as part of the main ontology because they are declared but never imported.

Finally, the third challenge involved implementing changes in our terminology server to support dealing with very large FHIR code systems. Our initial implementation assumed that large code systems would be imported using a specific component, such as the RF2 importer we use to support SNOMED CT, and that FHIR code systems would be small. Producing a FHIR representation of OWL ontologies instead of importing them directly has the advantage that the generated resources are standardised and can therefore be, in theory, uploaded to any FHIR terminology server. However, this also means that now FHIR code systems can be expected to contain thousands of concepts, as is the case with HPO.

3 Results

The transformation was implemented as an open source Spring Boot command line interface application. The OWL-API library was used to load the input OWL file and the ELK reasoner was used to classify the ontologies, because all

2The concepts are CHEBI_18183, CHEBI_17553, CHEBI_89843, CHEBI_17012, CHEBI_28875, CHEBI_132952, GO_0016265 and UBERON_0001029. The transformation was tested with HPO version 20180613.

3Available at https://github.com/aehrc/fhir-owl.
the ones selected for this project fall into the OWL EL profile\(^4\). The OWL-API library allows plugging in a different reasoner if required. The OWL ontologies are always classified before the transformation so this ensures that the is-a relationships available in the FHIR representation are inferred. The HAPI-FHIR libraries were used to generate the FHIR in-memory model and to save it to disk in JSON format.

HPO was used as an initial case study and the transformation produced a compliant FHIR code system that was loaded into Ontoserver and can be visualised correctly using Shrimp, our FHIR terminology browser. Figure 2 shows the HPO hierarchy displayed in Shrimp.

![Figure 2: The Human Phenotype Ontology in the Shrimp browser.](image)

To make sure that the proposed approach is generalisable to any OWL ontology, a set of additional ontologies that are relevant in the genomics space were selected and transformed into FHIR code systems. Table 3 shows the list of ontologies that were transformed.

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Version</th>
<th># of Classes</th>
<th>Parameters</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundation Model of Anatomy</td>
<td>4.12.0</td>
<td>104523</td>
<td>-name</td>
<td>The name is not in the OWL file. There are no imports but one external class is declared in the ontology, so the main namespace is required.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-mainNs</td>
<td>Some external concepts are declared in the ontology so the namespace is required to correctly identify local and imported concepts.</td>
</tr>
<tr>
<td>Human Ancestry Ontology</td>
<td>2.1.1</td>
<td>574</td>
<td>-mainNs</td>
<td>The name is not in the OWL file.</td>
</tr>
<tr>
<td>Human Phenotype Ontology</td>
<td>2018-06-13</td>
<td>32531</td>
<td>-name</td>
<td>The main namespace is required to correctly identify the local and imported concepts.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-mainNs</td>
<td>The name is in the OWL file, but in the dc:title annotation.</td>
</tr>
<tr>
<td>NCI Thesaurus</td>
<td>2018-09-04</td>
<td>138609</td>
<td>-nameProp</td>
<td></td>
</tr>
</tbody>
</table>

\(^4\)Strictly speaking, HPO contains axioms that fall outside of the EL profile, but because the Editor’s Guide for HPO developers, available at [https://github.com/obophenotype/human-phenotype-ontology/wiki(Editor-Guide)](https://github.com/obophenotype/human-phenotype-ontology/wiki(Editor-Guide), recommends using ELK, we assume that ignoring these axioms does not produce a significant difference in the classified form.
Table 3: List of OWL ontologies transformed into FHIR code systems.

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Date</th>
<th>ID</th>
<th>Column1</th>
<th>Column2</th>
<th>Column3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI Thesaurus Plus</td>
<td>2018-09-04</td>
<td>143186</td>
<td>-name</td>
<td>-name</td>
<td>The name is in the OWL file, in the dc:title annotation, but it is not NCIt Plus as expected, but rather NCIt OBO Edition.</td>
</tr>
<tr>
<td>Orphanet Rare Disease Ontology</td>
<td>2.6</td>
<td>13871</td>
<td>-name</td>
<td>The name is not in the OWL file.</td>
<td></td>
</tr>
<tr>
<td>Phenotypic Quality Ontology</td>
<td>2018-03-28</td>
<td>1737</td>
<td>-name</td>
<td>The name is not in the OWL file.</td>
<td></td>
</tr>
<tr>
<td>HUGO Gene Nomenclature</td>
<td>2010-07-12</td>
<td>32918</td>
<td>-name</td>
<td>-version</td>
<td>-url</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-code</td>
<td>-display</td>
<td>-synonyms</td>
</tr>
</tbody>
</table>

4 Discussion

The original motivation for this work was to provide access to HPO using the FHIR API but the implementation evolved to become a generic OWL to FHIR transformation engine. This enables access to many existing OWL ontologies through the FHIR API. One of the benefits of having access to this content is that it is now possible to replace existing terminology servers with FHIR-compliant ones without missing out on available content. For example, REDCap\(^5\) provides autocomplete-style search over OWL ontologies through BioPortal, which only provides basic searching functionality. A new REDCap plugin\(^5\) provides an alternative implementation that uses a FHIR server that delivers more powerful searching capabilities than BioPortal, such as constraining search to value sets, i.e., to a relevant subset of an ontology. For example, if a specific field is capturing a cancer type, a value set can be used to constrain the possible values to search for the children of the cancer concept in the target ontology, instead of having to search through all the content, which will most likely include a lot of irrelevant concepts. This is very useful when trying to improve the quality of data being captured, especially when dealing with large terminologies. The OWL to FHIR transformation allows importing all the OWL content in BioPortal to a FHIR server.

The solution addresses the challenges adequately but it does have some minor drawbacks that could be solved with changes or extensions to the main FHIR specification, to allow directly referencing parents and children across code systems. This would require the parent and children properties to be of type coding and all the operations that use subsumption would also need to be updated to support search parameters of type token. We are currently exploring these potential changes with the FHIR community and assessing the possible impacts in the current implementation of our terminology server.

One limitation of this work is that it assumes that the target FHIR concepts are only OWL classes. While this is true in all the ontologies that were transformed as part of this project, it might be the case that an ontology also contains individuals and these should also be transformed into FHIR concepts. Also, anonymous classes are not supported, because in FHIR a concept's code is a mandatory field. Finally, only a set of predefined properties are created in the target code system, mostly used to build the hierarchy correctly. Arbitrary properties in the source OWL file, represented as annotation properties in OWL, cannot be mapped to the target code system with the current implementation. Some OWL constructs, such as concrete domains\(^9\), could also be represented in some simple cases as properties in the target code system, but are currently ignored.

\(^5\)Available at [https://github.com/aehrc/redcap_fhir_ontology_provider](https://github.com/aehrc/redcap_fhir_ontology_provider).
5 Conclusion

In this paper we described the implementation of a transformation between OWL ontologies and FHIR code systems. The implementation maps elements using sensible defaults and also supports many configuration parameters.

A case study that is relevant in the genomics space was conducted using HPO and this highlighted some of the challenges in the transformation. Several approaches were considered and a satisfactory solution with minimal drawbacks was achieved. The solution was then tested on a larger group of ontologies and this showed that the transformation is generic and works with arbitrary OWL ontologies. In order to overcome some of the minor drawbacks, changes or extensions to the FHIR specification are currently being considered. These would allow natively referencing parents and children across code systems to better support importing external content.

The main benefit of this transformation is that it enables access to a broad range of ontologies in OWL format through the standard FHIR terminology APIs, without requiring the implementation of a bespoke importer for each one. This supports their use in electronic data capture systems such as EMRs through a single interface.

6 Acknowledgements

This work was partly funded by Australian Genomics. Australian Genomics is supported by the National Health and Medical Research Council (GNT1113531).

References


ITCBio, a Clinical and Translational Research Platform
Jesús Moreno-Conde, Alberto Moreno-Conde, Samuel Salas-Fernández, Carlos L. Parra-Calderón
1Group for Research and Innovation in Biomedical Informatics, Biomedical Engineering, and Health Economy. Institute of Biomedicine of Seville, IBiS/"Virgen del Rocío"
University Hospital /CSIC/University of Seville, Seville, Spain; 2Virgen Macarena University Hospital, Seville, Spain

Abstract
An informatics platform has been designed, deployed and validated around the ITCBio initiative to provide support to clinical and translational research in Andalusia. To this end, an infrastructure has been developed which, in a scalable manner, incorporates functionalities aimed to facilitate the consistent definition of information models, the data reusability from electronic health records, as well as the analysis and processing of information. All this with the purpose of providing support to the different clinical and translational research processes associated with clinical trials and research projects. This initiative is based on the creation of a suite of applications that, through using standards, incorporates open-source tools intended to support these research processes. It is currently in widespread and growing use in university hospitals in which the platform is deployed.

Introduction
Research and innovation from the discipline of biomedical informatics for clinical and translational research EHR-based faces several significant challenges. In this paper, we present the ITCBio platform that we have developed to respond to these challenges in practice. This initiative is well accepted among users due to the high capabilities offered to clinical and translational research, from the integration and improvement of a set of tools of proven effectiveness, as described in the section on methods.

Informatics oriented to Clinical Research
Currently, the development of research studies is carried out by collecting data on paper or using a computer system proposed by the promoter of the research. This working scheme turns the data collected in the Electronic Health Record (EHR) into systems that cannot be reused for research, resulting in a double collection of information by professionals. This double data collection allows errors and inconsistencies to appear in the clinical information that could be controlled if the documentation is performed in a unified manner.

To accelerate and optimize the translation of information and knowledge that occur in each phase of the clinical research process, a new discipline oriented towards Clinical Research (CR) has emerged in the field of Informatics. Such discipline proposes the development of new theories, tools and computer solutions that allow to improve the research capacity of the institutions that adopt them [1]. These tools and computer solutions must be designed to support the four phases of the clinical trial (Phase I, Phase II, Phase III and Phase IV) and the rest of the processes involved in each research study. That is why the result of the advances in CR range from the increase of the efficiency in the collection and acquisition of data to the optimization of the design of protocols and its efficient management. Likewise, the recruitment of patients and their management, the notification of adverse events, the compliance and adherence to the legal framework, as well as the storage, the transfer, the process and the analysis of research data.

Research networks based on the reuse of Electronic Health Record data
Promoters of research studies have a period of protection for their discoveries limited to 20 years for new drugs and patent molecules. Considering that the average time required by the pharmaceutical industry to introduce a new drug to the market from the start of its research is about 12 years [2], promoters of research studies will have significant benefits if the development times of these studies are reduced.

Based on the significant benefits of reducing clinical trial development time, the pharmaceutical industry promotes the reusability of EHR data to reduce patient recruitment times. At European level, the EHR4CR project [3] funded by the IMI call has developed a framework of services for more efficient clinical research. The project is developing
a framework in which interoperability solutions are integrated through the application of standards, security, data quality, data storage, organizational processes, accreditations and certifications that allow to develop pilots where studies are conducted to a European level. Associated with the results of this project, the EHR4CR Champion program has been created with the objective of creating a network and community of hospitals oriented towards the reusability of data for research. Numerous European hospitals are currently being contacted to participate in this network [4-5].

Among the benefits offered by participating in this network to healthcare providers, it is identified that by having optimized mechanisms for conducting clinical trial feasibility studies, network members are eligible to participate in a more significant number of clinical trials. The associate hospitals of this network accept the installation of a platform in their hospitals that is in charge of returning anonymized data from the patients of the center to the researchers of the pharmaceutical industry during the feasibility study phase of the clinical trial. The Swiss-based company Clinerion has also developed a system with a similar approach in which it offers to install a free system in hospitals that allows consultations by researchers from the pharmaceutical industry with the aim of encouraging the hospital to participate in a larger number of clinical trials [6].

Finally, the ITCBio platform aims to offer an alternative to avoid technological dependence on the pharmaceutical industry and to ensure the control of data through the development of independent studies that promote networking between research centers.

Security and data protection

The new European General Data Protection Regulation (GDPR (EU 2016/679) has emerged as a response to the need to ensure that security and data protection mechanisms are established in new scenarios [7]. GDPR is a step forward in clarifying how to specify patients' consents to the application of their data in research studies and also clarifies that data protection principles should not be applied to anonymous information. Thus, anonymized data allows the data subject not to be identifiable and therefore, it can be used for statistical or research purposes. Thanks to this new regulation there is finally legal certainty in areas that were not sufficiently identified in the previous regulation.

In parallel, the development of initiatives that tries to democratize and share the use of biomedical and health data through the FAIR principles guidelines is becoming popular. Such initiatives seek to establish guidelines so that resources, data and research results are easy to find, with clear rules of access, interoperable and reusable (“Findable, Accessible, Interoperable, and Reusable” [8]). The FAIR4Health project guides the future strategy of the European Commission concerning the management of publicly funded research data. This project is coordinated by the first author of this paper on behalf the Andalusian Health Service (SAS) and has 17 partners from eleven countries (Spain, Portugal, Italy, United Kingdom, Germany, Switzerland, Austria, Netherlands, Belgium, Serbia, and Turkey). Our ITCBio platform will be improved for the anonymization and implementation of the FAIR principles for the data collection that the Andalusian Health Service will bring to the FAIR4Health project.

Currently, the use of new technologies allows the establishment of new paradigms based on Dynamic Consent. Dynamic Consent takes advantage of information technologies to establish more flexible mechanisms of communication with patients and provides them greater control of their data and biological samples [9].

Interoperability

From the point of view of health care providers, due to the need to support the continuous care of patients in different health institutions, there are several initiatives currently working on the definition of specifications and computer models that allow communication between EHR systems and laboratory systems, pathological anatomy, radiology, etc.

To support the migratory flows and movements of people in Europe, the European Commission has defined a series of services focused on cross-border healthcare in our continent. These services are called Connecting Europe Facility and they are based on a European interoperability framework for eHealth that has defined the most relevant use cases in which there will be an emphasis at European level on the standardization of communications. The European interoperability framework defined identifies how to apply standards and profiles of HL7, IHE and Continua Health organizations for communication in telemonitoring scenarios, as well as the transfer of patient’s summary of records, prescriptions and drug dispensing, discharge reports, laboratories and radiology.
Besides, at national level, the Spanish Ministry of Health, Social Services and Equality has also recently defined the Minimum Data Set of Clinical Reports to support the continuity of patient care across multiple autonomous communities through the application of the HL7 CDA standard.

Moreover, the HL7 organization has recently developed the HL7 FHIR standard as a technology that enables the communication of clinical information [10]. This technology is having a great acceptance in a part of the industry, especially in the development of mobile applications, since it allows REST communications. There is a new initiative called SMART that establishes a framework for the development of mobile applications in the healthcare field using FHIR technology [11]. Based on these specifications, the Sync for Science (S4S) initiative has been established in the USA, with the participation of the leading EHR system commercialization companies (e.g., Allscripts, Cerner, eClinicalWorks, and Epic, etc.) and organizations such as the National Institutes of Health (NIH), Office of the National Coordinator for Health IT (ONC), and Harvard Medical School.

**Andalusian Health Service (SAS)**

Servicio Andaluz de Salud-SAS (Andalusian Health Service) is a public body providing all healthcare public services to the 8.4 million inhabitants in Andalusia. It is an autonomous organization, attached to the Regional Ministry of Health of the Government of Andalusia. SAS provides a wide range of healthcare services from preventive and public health activities to highly specialized services, in a system with universal coverage and funded by taxes, through a network of integrated healthcare facilities organized to ensure the accessibility of the population: 1,500 primary care centers, 49 hospitals, and over 100,000 employees.

SAS Central Service is located in Seville (Spain), including the Deputy Directorate of Information Technology and Communications (STIC), responsible for the definition and implementation of ICT policies and strategies throughout the Andalusian Health Service. The STIC is leading the process of transformation and improvement of eSystems, to facilitate and improve the daily work of public healthcare professionals and the services provided to citizens.

Diraya (“knowledge” in Arabic) is the region-wide EHR system network-based that aggregates all the health-related information as a Single Health Record. The system, integrated and accessible throughout the whole Andalusian Health Service, links patients’ information generated at any of the different care levels. The electronic health record is structured in a federated way, with some information blocks kept at a central level (such as problems, allergies, past medical history, prescriptions, lab, and imaging tests…) and local blocks that store reports (discharge summaries, care plans…) generated during local care encounters.

For research, SAS covers the performance of research and innovation activities. As part of this activity, according to data collected in 2017 [12], SAS has developed and participated in a total of 1100 projects, 49 of which correspond to international research initiatives. In terms of clinical trials, SAS carried out a total of 2,143 trials in 2017, of which 1,753 were commercial initiatives while 390 were promoted as independent trials.

In order to respond to the challenges of research in this context of Andalusia, our proposal has been to develop an infrastructure to support the translational and clinical research processes with the following features:

- Highly scalable. The infrastructure will be able to grow according to the available resources.
- Based on open-source software. The use of this kind of applications avoids dependencies with proprietary solutions, counting on a wide community of users and developers who give support and make this type of applications evolve.
- Oriented to the normalization, integration, advanced analysis and visualization of biomedical information. To support the complete process of clinical and translational research.
- Reuse of components and information resources, which makes it highly effective.

**Methods**

A suite of applications capable of integrating different open-source software initiatives to support clinical and translational research has been developed. Specifically, the current version of the platform includes the integration of the following tools:

- Data warehouse management software: i2b2 [13]
- OpenClinica [14] and RedCap [15] clinical trial and research study management software
- Advanced biomedical information analysis software: tranSMART [16]

For the integration of these initiatives with the ITCBio platform, the service bus known as Mirth Connect has been used, a free software engine developed specifically for health integrations and which natively provides the HL7 messaging standard. Mirth Connect consists of a set of tools that allows us to develop, test, deploy and monitor connection interfaces in a comfortable and intuitive way through the definition of integration channels.

Through this integration engine, it is also possible to interoperate with the clinical information present in the EHRs of the SAS for patients recruited in clinical trials and research projects. This integration is carried out through the development of specific channels that allow the on-demand reuse of patients’ clinical information in the electronic forms designed for clinical trials and research projects.

For user authentication in the ITCBio platform, as well as in the different open-source software initiatives mentioned above, the following considerations are taken into account:

- Use of corporate services for SAS user identification. The ITCBio platform makes use of connections with the SAS Active Directory server (LDAP) for the reuse of corporate credentials of SAS researchers.
- Ability to integrate with other sources of authentication external to the ITCBio platform.
- Use of an Active Directory (LDAP) server for local users to the platform who are not present in the external authentication systems discussed above.
- Mapping of user roles and permissions according to the Study/Center to which it belongs and its corresponding association in the different applications integrated within the ITCBio platform.
- After the initial authentication of the users in the ITCBio platform, the user can move freely between the different integrated applications, offering the capacity of Single Sign On within the whole platform.

The ITCBio platform implements a context management system for research. This initiative follows a scheme similar to the one proposed by the standard Clinical Context Object Workgroup (CCOW) [17-18], a Health Level Seven International standard protocol designed to enable disparate applications to synchronize in real time, and at the user-interface level. In addition, the ITCBio platform also incorporates an interoperability module based on IHE XDS [19] profiles for the management and exchange of documents with biomedical information, as well as the CDISC [20] standard, which allows the information obtained in the course of studies to be available in any other system that interoperate through these standards, widely used in the sector. On the other hand, the ITCBio platform also offers other interoperability interfaces to accept information from external systems in FHIR format. This interoperability interface makes it possible to integrate clinical information into the i2b2 clinical information repository, an initiative already pursued by other researchers in the sector [21].

Results

From a functional point of view, the ITCBio platform provides many advanced capabilities that support the main research processes present in our organization. Below there is an indicative scheme where the main aspects mentioned above can be observed:
The ITCBio platform is currently deployed in the Virgen del Rocío University Hospital, the Virgen Macarena University Hospital and the Biomedicine Institute of Seville (Seville, Spain). It is also being used in 27 research projects and six clinical trials, many of which are multicentric initiatives. These projects are framed within diverse clinical domains such as Oncology, Allergology, Thoracic Surgery, Pediatrics, Neurology, Pneumology, Gynecology or Hematology. Among all these projects, there are more than 4,500 registered patients in multiple research studies and more than 300 registered users.

Currently, we have an active innovation project around this initiative funded by the Ministry of Health of the Regional Government of Andalusia. This project aims to define, implement and evaluate a sustainable and scalable service to support clinical and translational research through the ITCBio platform. Likewise, this project intends to identify requirements to meet the needs in order to establish a sustainable service in SAS, which allows to implement and assess the impact of offering the service to support research through an evolved version of the platform. Based on the use of the service and its benefits obtained, the goal is to estimate the potential impact of the extension of this service in Andalusia and to define a strategy for the promotion of Health research activity in our region.

Discussion

One of the main milestones achieved through this platform has been to develop a common framework that integrates heterogeneous open-software initiatives to improve the processes associated with clinical and translational research. As expected, each of these initiatives uses completely different technologies and data models, which require a significant workload to achieve their correct integration. Although it has been identified diverse interoperability standards at different levels (information exchange, patients' identification or users' identification, among others) that could greatly facilitate this integration process, the various software initiatives prefer to opt for independent developments, practically on the margin of all of them. This fact emphasizes that the current development of open-software tools related to the biomedical environment is not sufficiently aligned with the specifications defined by the standardization organizations. Consequently, the development of cross-sectional solutions, such as the ITCBio platform, is challenging.
Around this situation, we have found some standards that have entirely fallen into disuse, such as the standard of exchange of clinical context information (CCOW) and others that, despite their usefulness, are not yet on the roadmap of the open-source initiatives that we have thought about in this document. On the other hand, in order to achieve interoperability between systems, the most appropriate standard for communicating information about clinical trials and research studies is the CDISC ODM. This standard allows exchanging not only the information contained within the clinical trial or research study but also a large amount of metadata associated with its research protocol. Unfortunately, the open-source initiatives oriented to data exploitation (i2b2 and tranSMART) do not have this capacity yet, so in this project, it has been necessary to implement integration mechanisms with greater complexity that directly require working at the database level.

Besides, another of the main aspects to be highlighted in the development process of this platform is the modeling and management of clinical information recorded in clinical trials and research studies. Within this process, the ITCBio platform has developed a specific module oriented to the design of the electronic data forms by using semantic mechanisms. This module, based on a previous project known as HEMIC [22], is a tool that allows the consistent definition of forms based on the main reference terminologies and ontologies in the health domain. Through this module, the integration of the different open-source tools that are part of the ITCBio infrastructure has been facilitated. Besides, it has been possible to significantly improve the processes of clinical information exchange from the registration tools to the tools in charge of the analysis and data exploitation. In this way, we believe that the incorporation of semantic techniques for the definition of electronic data forms can be very useful for this kind of open-source initiatives, such as OpenClinica and RedCap.

Conclusion

The ITCBio platform begins with competitive funding of the call for Scientific-Technical Equipment of the Spanish Ministry of Economy and Competitiveness in 2013. The design, development, and deployment of a scalable and reusable infrastructure to support translational and clinical research based on open-source software were carried out. Its objectives are the following:

- To facilitate feasibility studies and patient recruitment in clinical trials based on Real World Data.
- To provide researchers with access and management of all the information they need to carry out their research activity, including access to sources of available clinical care information.
- To promote collaboration between researchers inside and outside each organization through tools that support the management of roles and accesses in multicenter studies, complying with regulations, security information and confidentiality needs, as well as the enrichment of data collections with metadata in accordance with FAIR principles for effective reusability of the same in new collaborative research.
- To offer tools for the advanced analysis and exploitation of the data obtained in the research studies.
- To develop new lines of research on the standardization, integration, advanced analysis and visualization of integrated information with available clinical information.

The ITCBio platform is based on the integration of different modules or open source components. Thanks to this integration of components, the platform can offer a set of services to support translational research that is not able to offer the components individually or any of the competitors in the market. In particular, this platform offers the following benefits:

- At the scientific level, the ITC-Bio platform provides advanced capabilities for the management and conduct of clinical trials and research studies with cross-cutting application to the whole range of diseases and pathologies. The ITC-Bio platform accelerates this discovery process and therefore has a great impact on the research and development of new treatments, making the health and research centers that use its services more attractive to participate in clinical trials promoted by the pharmaceutical industry, enhancing the capacity to generate research results that can be transferred to the biotechnology market.
- From the scientific-academic point of view, this infrastructure boosts the scientific production capacities of translational researchers and clinicians who make use of it, also boosting the development of clinical trials and research studies.
From the technological point of view, the ITC-Bio platform presents an important advance in terms of research platforms thanks to the inclusion of functionalities that allow collaboration between users and different institutions through the possibility of integrating with multiple user management systems. The platform does not depend on any external license because it integrates different open source applications in a common framework using a Service-Oriented Architecture that allows its scalability.

On the other hand, this infrastructure will be able to support all lines of translational and clinical research that may arise in the hospital environment. Also, its deployment and implementation are leading to the emergence of new lines of research in technologies applied to health and information technology for biomedical research, among others:

- Research and development of new interfaces for the efficient and effective visualization of information and decision support systems based on personalized medicine.
- Research and development of algorithms and advanced information analysis methodologies including the application of artificial intelligence techniques such as data mining and clustering.
- Research in rapid phenotyping techniques based on the application of natural language processing algorithms on clinical information stored in free text.
- Research on the semantic representation of biomedical information and the needs of the patient.
- Finally, it is intended to incorporate FAIRification capability (provision of FAIR-compliant metadata to research data collections) that would allow universal reusability of data collections which extends the interest in the use of the platform to all actors willing to open science and the acceleration of shared biomedical research based on open data.

Acknowledgement

Thanks to MD Esther Román Villarán for the editorial support to this manuscript. The present study was partially funded through the Ministry of Health of the Government of Andalusia in the ITCBio project (Project Reference PIN-0441-2017), and partially funded through FAIR4Health project, from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No 824666. On the other hand, this project has also been promoted by the Platform for Innovation in Medical Technologies and Health (Plataforma ITEMAS, code PT17/0005/0032) from “Carlos III” Spanish National Institute of Health.

References

11. S4S Initiative Webpage http://syncfor.science/
13. i2b2 Webpage https://www.i2b2.org/
15. RedCap Webpage https://www.project-redcap.org/
Unleashing the value of Common Data Elements through the CEDAR Workbench

Martin J. O’Connor, MSc¹, Denise B. Warzel, MSc², Marcos Martínez-Romero, PhD¹, Josef Hardi¹, Debra Willrett, MS¹, Attila L. Egyedi¹, Aras Eftekhar³, John Graybeal¹, and Mark A. Musen, MD, PhD¹

¹Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
²Cancer Informatics Branch, National Cancer Institute, Bethesda, MD, USA
³Attain LLC, McLean, VA, USA

Abstract

Developing promising treatments in biomedicine often requires aggregation and analysis of data from disparate sources across the healthcare and research spectrum. To facilitate these approaches, there is a growing focus on supporting interoperation of datasets by standardizing data-capture and reporting requirements. Common Data Elements (CDEs)—precise specifications of questions and the set of allowable answers to each question—are increasingly being adopted to help meet these standardization goals. While CDEs can provide a strong conceptual foundation for interoperation, there are no widely recognized serialization or interchange formats to describe and exchange their definitions. As a result, CDEs defined in one system cannot be easily reused by other systems. An additional problem is that current CDE-based systems tend to be rather heavyweight and cannot be easily adopted and used by third-parties. To address these problems, we developed extensions to a metadata management system called the CEDAR Workbench to provide a platform to simplify the creation, exchange, and use of CDEs. We show how the resulting system allows users to quickly define and share CDEs and to immediately use these CDEs to build and deploy Web-based forms to acquire conforming metadata. We also show how we incorporated a large CDE library from the National Cancer Institute’s caDSR system and made these CDEs publicly available for general use.

Introduction

The use of ontologies and controlled terminologies has become pervasive in biomedicine. Dozens of large ontologies and terminologies and hundreds of specialized smaller ones have been developed to cover many biomedical domains. For example, the National Center for Biomedical Ontology's BioPortal ontology repository¹ serves over 700 biomedical ontologies and terminologies, which are used throughout biomedicine. While ontologies and controlled terminologies provide a common vocabulary to refer to biomedical concepts, additional detailed specifications are typically needed to satisfy data collection and reporting needs.² Data collection requirements for clinical studies, for example, mandate precise specifications of questions and the range of possible answers to those questions. Common data elements (CDEs), which provide the means to link the specification of the question to a range of possible answers, are increasingly being adopted to satisfy these data collection and reporting needs.²⁻⁴

A CDE is effectively an agreed-upon question specification, precisely defining how a particular question should be asked and what values should be presented to users for their selection.⁵⁻⁶ An answer is represented as a single typed value. Answers can be strings, in which case simple parameters such as string length and encoding can be specified. They may be numeric, in which case parameters such as valid ranges, units, and precision may be specified. Answers can also come from controlled terminologies, where the answer may be specified as a terminology code or an encoded value. Often a permitted value set, or value domain, is built to define the allowed answers for a question. Such value sets are often built from one or more standards or terminologies.⁷⁻⁸ CDEs are commonly used to define case report forms (CRFs) for clinical trials, though they can be used in any situation where it is important to meet rigorous data collection or reporting requirements. They have been adopted most widely in cancer research,⁹⁻¹² but they are also used in other domains, such as epilepsy,¹³ brain injury,¹⁴ stroke,¹⁵ phenotyping,¹⁶ and radiology.¹⁷

CDE definitions can provide a strong foundation for interoperation because they allow data descriptions to be recorded in a registry. Such registries can help to standardize the way data are collected, stored, transferred, and reported. One of the largest CDE registries has been developed by the U.S. National Cancer Institute (NCI) with the goal of facilitating multidisciplinary, multi-institutional cancer research. This registry is called Cancer Data Standards Repository (caDSR)⁹,¹⁸ and it contains over 60,000 CDEs that cover many aspects of cancer research. The U.S.
National Institutes of Health (NIH) are also developing a multi-discipline registry that aims to unify the range of biomedical CDEs that have been produced by a variety of NIH and other organizations (https://cde.nlm.nih.gov).

While these registries provide a strong conceptual foundation for the definition and use of CDEs, there are a number of practical challenges that must be addressed when attempting to reuse CDEs. The common structure is provided by the ISO/IEC 11179 standard,19 which is used as a basis of the caDSR and other repositories, is helpful but the standard does not specify implementation-level details. As a result, CDEs cannot be easily used across systems. Reflecting the strong regulatory requirements of the domains they are used in, and ISO standard conformance requirements, building new CDEs tends also to be a laborious task that involves complex workflows. In addition, CDEs must be fully specified before data collection forms can be built, adding additional complexity. Building data collection forms from CDEs can thus be an onerous task. An additional issue is that ISO/IEC 11179-based systems typically aim to provide faithful implementations of large parts of the standard, significantly increasing their complexity.

There is thus a technological barrier to developing and reusing CDEs in new systems that we believe is limiting their adoption by the broader biomedical community. To address this problem we extended an existing Web-based metadata management platform called the CEDAR Workbench20,21 to provide a core representation of CDEs suitable for specifying questions in a metadata acquisition system. Rather than aiming to provide CDE definitions that reflect a comprehensive implementation of the ISO/IEC 11179 standard, we instead concentrated on providing the functionality in core parts of the standard that relates to the precise specification of questions and the values used to answer those questions. A key focus is on interoperation with Linked Open Data by providing a direct mapping of CDE-described data to RDF. We describe the functionality provided by the system and show how the resulting system allows users to easily use CDEs to build and deploy Web-based forms to acquire conforming data. We also show how we incorporated the large CDE library from NCI's ISO/IEC 11179-based caDSR system and made these CDEs publicly available for general use.

Methods

Although there are no widely adopted standards defining CDEs in the biomedical domain, the ISO/IEC 11179 specification19 has been used as the underpinning of many CDE-based systems. This specification, which is formally known as the ISO/IEC 11179 Information-Technology Metadata Registry standard, is divided into six parts and covers a wide range of requirements for developing and deploying metadata registries. Part 3 of the standard describes a Data Element, which is the fundamental information component in the standard. An ISO data element is designed to support the description of an atomic piece of data. It has been adopted by many systems to model common data elements.

In addition to defining the core specification of a particular question and answer, the standard also outlines a rich model describing many aspects of a data element. Such information includes provenance information (e.g., who developed the data element, workflow (e.g., its development status, such as whether it is under development or is released), possible relationships with other data elements, and detailed descriptions about the context and the domain in which the data element is to be used. While useful in some situations, this information is generally not needed if the goal is to produce an operational implementation of a system that uses CDE-based question specifications. In this paper we, ignore these contextual metadata. Instead, our analysis and implementation restrict themselves to the core features needed to deploy CDEs in a metadata acquisition system.

To provide a detailed set of requirements for supporting the resulting types of CDEs, we analyzed a 11179-based CDE system. In particular, we used the cancer Data Standards Repository (caDSR)22,23, which was developed by the U.S. National Cancer Institute. The system was designed to support the development and deployment of CDEs in data collection forms in cancer research. It adheres very closely to the ISO/IEC 11179 Edition 2 metadata standard, with extensions to support terminologies and ontologies by data elements. It provides a rich and comprehensive implementation of that standard and has been used for over two decades by clinical-trials data management systems.

Requirements

Using the caDSR system as the source, our requirements analysis aimed to identify the core set of features necessary to support CDEs in a metadata acquisition system. In addition to faithfully representing the core requirements of a CDE itself, a system must be able to ingest libraries of CDEs, be able to use them to build Web-based forms, and, finally, to be able to deploy those forms to acquire data meeting the specifications of the CDEs used in the forms. An additional requirement is to produce an RDF representation of data collected using CDEs.

The analysis identified three main sets of requirements for representing CDEs: (1) Descriptive information about a CDE must be represented. This information includes a public identifier for a CDE, its name, and a definition. CDEs
can be versioned, allowing them to evolve over time. CDEs can also be tagged with basic provenance information, which can be used to, for example, indicate their domain or to divide them into categories. (2) CDE question specifications must be represented. The caDSR system provides a number of ways to indicate the questions presented to users when they are filling in CDE-specified fields. In addition to the primary question text, CDEs may also contain an set of alternative questions. These alternatives questions can be used to customize deployed CDEs for different domains. (3) Finally, CDE value domains must be represented. An array of complex features is provided by caDSR to define the value domain of a CDE. Broadly speaking, values may be string-based, dates, times, Boolean, numeric, or from a controlled term source. String-based CDEs support fairly simple minimum and maximum length specifications. Numeric values may come from a variety of different datatypes. Units and ranges may also be specified for numeric values. Controlled term value domains specifications are the most complex. A system must allow value sets to be built to define sets of controlled values. These value sets may be customized during deployment to allow reordering of values so that the most useful values are presented to users for particular needs. Similarly, some values from a value set can be excluded for particular deployments. Again, these features are to allow customization of CDEs to meet different deployment needs.

A final set of requirements that we identified relates to the management of CDEs in a system. Users must first be able to search for CDEs for use in forms. Searching may be by CDE public identifier, name, description, category, and value in value set. Once a CDE is found, users must be able to add the selected CDE to a form, and possibly to customize the CDE. Customizations include selecting a question from a set of alternative questions and reordering or excluding values in a value set. Web-based forms must then be generated from these form specifications to acquire data from end-users. The deployed forms must enforce the value domain restrictions specified by CDEs.

**Implementation**

We decided to provide CDE support on top of an existing platform (Figure 1), rather than to implement the required features from scratch. This platform, which was developed by the Center for Expanded Data Annotation and Retrieval, is called the CEDAR Workbench. It provides a collaborative, Web-based environment for managing metadata resources. The platform is centered on the creation of metadata templates (or simply templates) to describe biomedical experiments. These templates define the data attributes—termed template fields or fields—needed to precisely describe these experiments. For example, an experiment template may have an organism field containing the name of the organism being studied by the experiment (e.g., Homo sapiens).

![Figure 1. High-level overview of the workflow of ingesting libraries of CDEs into CEDAR. CDEs from an external library are transformed to the CEDAR model and uploaded to the CEDAR Workbench via the CEDAR REST APIs. The controlled term value sets used by these CDEs are stored in BioPortal. The biomedical community can easily access and reuse the CDEs when building Web-based metadata acquisition forms. The CDEs can also be accessed via the CEDAR REST APIs.](https://example.com/figure1.png)
numbers, and dates, field values can also be defined using standardized terms from ontologies and other sources of controlled terms. CEDAR works in concert with the BioPortal ontology repository⁴ to supply these values. BioPortal is a popular platform for accessing and sharing biomedical ontologies and hosts over 700 ontologies and 8.9 million classes. This combination of CEDAR and BioPortal provides the ability to create value set descriptions that nicely align with the requirements of CDE value domains.

A final set of features provided by CEDAR supplies the ability to deploy CDEs in Web-based forms defined by templates. These are: (1) a Template Designer, which supports interactive template creation; (2) a Metadata Editor, which allows end-users to fill in templates with metadata; and (3) a Metadata Repository, which manages the storage and retrieval of both templates and the metadata created using those templates. CEDAR defines a standardized metadata model, together with Web-based services to store, search, and share metadata.²⁴ This model is based on the JSON Schema and JSON-LD specifications. It allows users to publish their metadata as both JSON-LD and RDF, thus facilitating interoperability with Linked Open Data. Users can quickly create forms using the Template Designer and deploy the forms using Metadata Editor to produce semantically annotated data.

CEDAR thus provides a set of core features that are suited to supporting the deployment and use of CDEs. Based on the requirements outlined earlier, we identified the set of existing CEDAR features and the set of extensions that would meet these requirements. The primary extensions involved significantly enhancing CEDAR's field-level capabilities to support the rich value restrictions specified by CDEs. In addition to this field-level functionality, CEDAR must also be able to ingest a CDE library and map the library’s CDE representation to the CEDAR model. With this functionality, CDE-based fields can be added to templates and deployed to collect metadata from users using CEDAR’s Template Designer and Metadata Editor tools.

**Field Extensions** We performed a detailed analysis of the caDSR implementation of ISO/IEC 11179 standard to identify the core functionality that a CDE system must support. The first task was to find commonalities between the caDSR model and the CEDAR metadata model, as well as the limitations of the latter to support CDE-based fields. Note that, reflecting the ISO/IEC 11179 standard, caDSR stores rich provenance information and domain-level semantic descriptions for individual CDEs. Since we are aiming only to represent the relevant operational specifications of CDEs, we excluded these from our analysis. We identified four classes in the caDSR model that contain information that represents this operational information. The classes are *Data Element*, *Value Domain*, *Permissible Values*, and *Reference Document* (Figure 2).

**Figure 2.** Visual representation of the mappings between the core caDSR CDE model and the CEDAR and BioPortal models. The figure shows the subset of attributes of the caDSR model that are mapped to the CEDAR Template Model (dashed blue line) and to the BioPortal Value Set model (dashed red line). The attributes `publicId`, `version`, `preferredName`, and `preferredDefinition` from the *ValueDomain* caDSR entity are mapped both to CEDAR and BioPortal.
The core CDE entity in caDSR model is the *Data Element*, which represents the smallest unit of data that can be represented and exchanged between systems. It contains attributes that capture descriptive information for a CDE, such as its public identifier, name, and version. The characteristics of the values accepted by a data element are defined by the *ValueDomain* entity, which contains attributes to specify the data type and the different value constraints. Each data element and value domain can contain terminology references. The attribute *ValueDomainType* of the *ValueDomain* entity specifies whether the domain is enumerated—when it is specified by a predetermined list of permissible values (e.g., *male* and *female* for the element *Sex*)—or non-enumerated—when the domain is specified by a description or range (e.g., positive integers for the element *Number of Months Stayed Off Cigarettes*). When the value domain is enumerated, the accepted values are represented by the *PermissibleValue* entity, which specifies the exact names, codes, and textual labels that can be stored for the CDE.

**Table 1.** Field-level mappings between the core caDSR model and the CEDAR and BioPortal models. The table shows the core attributes of a CDE in the caDSR model and the corresponding attributes in a CEDAR model. For example, the first row shows that the attribute *PUBLICID* from the element *Data Element* in the NCI’s caDSR model file is mapped to the field *schema:identifier* in the CEDAR model. The right column shows the attributes used to represent value sets and their values in BioPortal. Fields added to the CEDAR model to support the representation of CDEs are indicated by an asterisk.

<table>
<thead>
<tr>
<th>NCI caDSR CDE (XML)</th>
<th>CEDAR field (JSON Schema, JSON-LD)</th>
<th>BioPortal caDSR Value Sets (JSON)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataElement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PUBLICID</td>
<td>schema:identifier*</td>
<td>-</td>
</tr>
<tr>
<td>LONGNAME</td>
<td>schema:name, title, description</td>
<td>-</td>
</tr>
<tr>
<td>PREFERENCEDEFINITION</td>
<td>schema:description</td>
<td>-</td>
</tr>
<tr>
<td>VERSION</td>
<td>pav:version</td>
<td>-</td>
</tr>
<tr>
<td>WORKEFLOWSTATUS</td>
<td>bibo:status</td>
<td>-</td>
</tr>
<tr>
<td>ReferenceDocument</td>
<td>DocumentText</td>
<td>-</td>
</tr>
<tr>
<td>DocumentType</td>
<td>used to map DocumentText</td>
<td>-</td>
</tr>
<tr>
<td>DocumentText</td>
<td>skos:prefLabel</td>
<td>-</td>
</tr>
<tr>
<td>ValueDomain</td>
<td>PublicId, Version</td>
<td>ID, prefLabel</td>
</tr>
<tr>
<td>PublicId</td>
<td>-</td>
<td>identifier</td>
</tr>
<tr>
<td>Version</td>
<td>-</td>
<td>hasVersion</td>
</tr>
<tr>
<td>PreferredName</td>
<td>_valueConstraints.valueSets.name</td>
<td>altLabel</td>
</tr>
<tr>
<td>PreferredDefinition</td>
<td>-</td>
<td>comment</td>
</tr>
<tr>
<td>DataDomain</td>
<td>_ui.inputType</td>
<td>-</td>
</tr>
<tr>
<td>Datatype</td>
<td>_valueConstraints.numberType*</td>
<td>-</td>
</tr>
<tr>
<td>DecimalPlace</td>
<td>_valueConstraints.decimalPlace*</td>
<td>-</td>
</tr>
<tr>
<td>MinimumValue</td>
<td>_valueConstraints.minValue*</td>
<td>-</td>
</tr>
<tr>
<td>MaximumValue</td>
<td>_valueConstraints.maxValue*</td>
<td>-</td>
</tr>
<tr>
<td>UnitOfMeasure</td>
<td>_valueConstraints.unitOfMeasure*</td>
<td>-</td>
</tr>
<tr>
<td>MinimumLength</td>
<td>_valueConstraints.minLength*</td>
<td>-</td>
</tr>
<tr>
<td>MinimumLength</td>
<td>_valueConstraints.maxLength*</td>
<td>-</td>
</tr>
<tr>
<td>PermissibleValue</td>
<td>VMPUBLICID, VMVERSION</td>
<td>ID, prefLabel</td>
</tr>
<tr>
<td>VMPUBLICID</td>
<td>-</td>
<td>identifier</td>
</tr>
<tr>
<td>VMVERSION</td>
<td>-</td>
<td>hasVersion</td>
</tr>
<tr>
<td>VALIDVALUE</td>
<td>skos:notation*</td>
<td>notation</td>
</tr>
<tr>
<td>VALUEMEANING</td>
<td>rdfs:label</td>
<td>prefLabel</td>
</tr>
<tr>
<td>MEANINGDESCRIPTION</td>
<td>-</td>
<td>comment</td>
</tr>
<tr>
<td>MEANINGCONCEPTS</td>
<td>@id</td>
<td>relatedMatch</td>
</tr>
<tr>
<td>PBEGINDATE</td>
<td>-</td>
<td>startTime</td>
</tr>
<tr>
<td>PVENDDATE</td>
<td>-</td>
<td>endTime</td>
</tr>
</tbody>
</table>

When comparing the caDSR and CEDAR models, we noticed a direct correspondence between the caDSR *DataElement*, *ReferenceDocument*, and *ValueDomain* entities and the CEDAR *TemplateField* entity. We observed that the *TemplateField* entity contained attributes to cover some of the core information for CDEs, such as name, definition, and version. However, it lacked support for some other crucial CDE features. We identified five main limitations in the CEDAR model and associated software and developed functionality to provide them:

**Public Identifiers** The CEDAR model lacked a way to store a public identifier of a field. This information complements the field CEDAR identifier and it is crucial to accurately find and refer to the CDE. As a solution to this limitation, we added the *schema:identifier* property to the model, which can now be used to store public identifiers for any CEDAR field. The Template Designer and Metadata Editor were also extended to handle this new field.
The caDSR ReferenceDocument entity makes it possible to specify, in addition to the preferred question text, alternate questions that can be presented to the user when filling out CDE-specified fields. An example of preferred question text for a CDE is Has the disease relapsed? while an alternate question is Was the status considered a disease relapse? We extended CEDAR to allow entering an alternate question text and stored it in the model using the property skos:altLabel. The Template Designer and Metadata Editor were also extended to work with alternate questions.

**Datatypes** The caDSR ValueDomain entity contains a dataType attribute that specifies the type of the value accepted by the CDE. Our analysis of a set of 60,409 caDSR CDEs revealed 190 different data types. Most of these datatypes can be directly mapped to equivalent data types supported by CEDAR. To limit the implementation effort, we decided to initially map the 10 most used datatypes (CHARACTER, java.lang.String, ALPHANUMERIC, ISO21090CDv1.0, NUMBER, java.lang.Long, java.lang.Integer, java.lang.Double, java.util.Date, DATE). These 10 types are present in 53,175 CDEs (88%). We are currently working on mappings the remaining data types to existing CEDAR types. In a few cases, we will need to extend CEDAR to support less common data types.

**Value Constraints** The ValueDomain entity contains a rich set of attributes to define some advanced value constraints, which the CEDAR model did not support. We extended the CEDAR model with fields to specify the minimum and maximum values, the number of decimal places, and the unit of measure accepted by numeric fields. We also the added fields to store the minimum and maximum length of string fields. The Template Designer and Metadata Editors were also enhanced to handle these value constraint extensions.

**Value Sets** In caDSR value sets can be used to define the range of possible values for a CDE. Value sets are versioned, first-class entities. They can be reused by several CDEs and can evolve over time. Typically, the values in a value set are selected from controlled term sources. Most value set values in caDSR come from the National Cancer Institute Thesaurus (NCIT). In collaboration with the BioPortal team, we extended BioPortal to support the representation of caDSR value sets. The enumerated value set specified by a ValueDomain is mapped to a BioPortal value set, while all the permissible values are stored as values in the value set. In CEDAR, the CDEs with enumerated values are linked to BioPortal using a URI that identifies the value set in BioPortal. We also extended the CEDAR model to support value set reordering and value exclusions. The Metadata Editor was also modified to handle this reordering and exclusion functionality.

![Figure 3](image-url). Schematic showing the caDSR CDE ingestion workflow. The cadsr2cedar tool takes an XML file with a set of caDSR CDEs and transforms each CDE to a CEDAR field. The transformed CDE fields are uploaded to the CEDAR Workbench via the CEDAR REST API. Any associated value sets and their values are transformed into an OWL and then uploaded to BioPortal via the BioPortal REST API.

**Ingestion Pipeline** We developed a pipeline to ingest the set of public caDSR CDEs into the CEDAR Workbench. The process converts XML-encoded caDSR CDEs to JSON Schema-encoded fields in the CEDAR model (Figure 3). We used a set of 60,409 CDEs that we downloaded from caDSR on August 2018. We developed a tool called cadsr2cedar to transform the CDEs in the XML file to the CEDAR model. Table 1 shows the primary mappings used.
in this conversion. The conversion process excluded currently unsupported datatypes and also excluded CDEs that were not marked as released by the caDSR system. As a result, we obtained 48,817 CDEs, which we uploaded to the CEDAR Workbench using its REST API. The cadsr2cedar tool also generated a total of 8,808 different value sets and 114,121 values used by the CDEs. We uploaded these to BioPortal using the BioPortal REST API.

Currently, we run the conversion process manually, but we plan to enhance it to run to immediately reflect CDE updates in the caDSR system. Extensions will be needed to the caDSR system to notify 3rd-party systems of such updates. The ingestion tools will also need to quickly identify the CDEs or value sets that have changed since the last update and selectively update only the corresponding CEDAR or BioPortal entities.

All the CDEs ingested are publicly available in the CEDAR Workbench (https://goo.gl/SggjQB).

Results

The main result of this work is a new version of the CEDAR Workbench that supports CDEs and that has been equipped with a large library of CDEs defined and maintained by the NCI. In addition to the field types natively available in the CEDAR Workbench (e.g., text, date, numeric, ontology values), CEDAR users can now search and select from a large number of CDEs ingested from the NCI's caDSR registry to build Web-based data acquisition forms. These forms can be used to collect data based on standard values from ontologies and terminologies, and easily share both the forms and the collected data with the broader biomedical community.

We used a CDE ingestion pipeline to incorporate 49,280 NCI caDSR CDEs into the CEDAR Workbench, as well as to upload the corresponding value sets to the BioPortal ontology repository. Our CDE ingestion pipeline can be reused to ingest new NCI caDSR CDEs into CEDAR. It can also be adapted to include CDEs from other sources.

**Figure 4.** Parallel representation of a CDE in the caDSR and CEDAR systems. The left side of the figure shows a screenshot of NCI's CDE Browser with details of a CDE, including its public identifier (3111302), version (2.0), and long name (Specimen Type Collection Biospecimen Type). The right side of the figure shows how that particular CDE can be used in the CEDAR Template Designer to build a Biospecimen template in combination with other non-CDE fields (Study Title, Organism, Cell Type, and Test Date), which were created on-the-fly. The Template Designer displays the most relevant information for the CDE, including its version, public identifier, long name, preferred question text, definition, and value set. This value set can be explored interactively.

CDEs can now be used when building CEDAR templates (Figure 4). The left side of the figure shows a screenshot of the NCI's CDE Browser (https://cdebrowser.nci.nih.gov/cdebrowserClient/cdeBrowser.html) for the CDE Specimen Type Collection Biospecimen Type. The CDE Browser provides relevant details for the selected CDE, including its identifier (3111302), version (2.0), and the preferred question text that should be used when incorporating it into a form (Sample Type). The right side of the figure shows how that CDE can be used in the CEDAR Template Designer to build a Biospecimen template, in combination with either other CDEs or with other fields that can be created on-the-fly, such as Study Title and Organism.
Figure 5 shows a screenshot of the Metadata Editor for the Biospecimen template, with the list of allowed values for the Sample Type. The right side of the figure shows the entered values in JSON-LD format. For the Sample Type field, CEDAR’s metadata contain not only the label of the value selected (Blood) but also the URI of the corresponding term in the NCI Thesaurus (http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C12434).

Our current implementation allowed us to transform 88% of the set of 60,409 caDSR CDEs available in August 2018. We are currently working on extending CEDAR to reach 100% coverage.

Figure 5. Screenshot of CEDAR’s Metadata Editor displaying a data acquisition form generated from the Biospecimen template shown in Figure 4. Here, the user is about to select the value Blood from the list of allowed values for the Sample Type field. The right side of the figure shows CEDAR’s JSON-LD representation for the entered values. For the Sample Type field, the JSON-LD representation contains not only the label of the selected value (Blood), but also the URI of that term in the NCI Thesaurus (http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C12434). CEDAR’s JSON-LD metadata also contains the values entered for the other fields, as well as provenance information such as the author and the date when the values were entered. This JSON-LD representation can be automatically transformed to RDF.

Discussion

Over the past few decades, there has been a strong emphasis on standardizing data collection and reporting requirements to facilitate data discovery, data interpretation, and data reuse. Different communities have adopted a variety of approaches to address these standardization needs. The cancer community, for example, has adopted CDEs as a mechanism for defining reusable question specifications that can be employed when collecting and reporting data in clinical research studies. Large libraries of CDEs have been built, which provide a strong foundation for interoperability. The social sciences community often uses data collection tools such as REDCap for their reporting needs. REDCap allows users to construct data acquisition forms. These forms can be built from reusable data collection instruments, which are used to standardize data-collection requirements over different studies. In the biological sciences, controlled terminologies are widely used. Many comprehensive ontologies have been developed, providing a common vocabulary to refer to biological entities.

While these solutions address the needs of their respective communities, they do not interoperate with each other, and each has shortcomings. For example, approaches such a REDCap make little use of controlled terms. The CDE-based solutions developed by the cancer community typically involve systems that are not easily reusable outside the specialized task of constructing case report forms. And, while controlled terms are useful in themselves, they do not address the full needs of specifying data-collection and reporting requirements. There is a need for an interoperable approach that meets the needs of these different communities.

We believe that the CDE-based approach outlined in this paper can provide such a solution. The system we developed supports the easy form construction capabilities of tools such as REDCap and the precise data specification advantages of CDEs, together with the central use of controlled terms. The system is built using the principled definition of CDEs provided by the ISO/IEC 11179 standard. However, the goal was not to exhaustively represent the wide array of provenance information and conceptual metadata that is specified by the standard; such information would be needed,
for example, when developing a specification-conforming metadata registry, or when trying to identify or analyze CDEs or data that are semantically related. Instead, the system restricts itself to the core parts of the standard necessary to represent standalone question specifications. This restricted interpretation of the ISO/IEC 11179 standard leverages its power while simplifying the specification of robust CDEs. We have done some preliminary testing to demonstrate the feasibility of our approach with NCI users, though more empirical testing is required to fully evaluate our claim.

Conclusion

In this paper, we describe how we extended the CEDAR Workbench to natively support CDE-based question specifications. The primary goal of this work is to provide an open platform that dramatically simplifies the use and deployment of CDEs by supplying intuitive and highly interactive Web-based interfaces. A key focus is to support interoperation, both by allowing third-party CDEs to be incorporated into the system, and by representing CDEs using ontologies and Semantic Web standards. Support for this markup was provided by extending the BioPortal ontology repository to natively facilitate the creation of value sets of controlled terms and then allowing these values sets to be interactively linked to CDE definitions. The resulting functionality supports the creation of robust, semantically rich CDE definitions that can be quickly deployed to collect data.

We ingested a library of over 48,000 CDEs and associated value sets from the NCI's caDSR CDE repository and made these CDEs available for public use. In particular, we validated that the system could present questions to users that accurately reflect a CDE specification and that it could also ensure that acquired answers fully meet the value requirements of those specifications. As outlined in this paper, we made several extensions to CEDAR to support caDSR CDEs, eventually reaching 88% coverage. We are currently developing additional features to fully represent the remaining caDSR CDEs. In addition to making the CDEs available for reuse, the goal of this ingestion task was to validate that CEDAR could faithfully represent and enforce a subset of the key elements of an ISO/IEC 11179-based CDE specification. This initial set of CDEs was developed over several decades for use in a large number of case report forms for clinical trials. They cover a broad range of cancer research. We plan to ingest CDEs from a variety of other sources to increase domain coverage. We also plan to develop additional features to support the management and search of multiple, large CDE collections.

The resulting system provides an open platform for sharing, managing, and deploying CDEs. We believe that the system lowers the barrier to the use of CDEs by leveraging existing CDE libraries and by supporting the easy creation of data-collection forms that allow these CDEs to be quickly provided to end-users.

Acknowledgements

CEDAR is supported by the National Institutes of Health through an NIH Big Data to Knowledge program under grant 1U54AI117925 and partially sponsored by NCI/FNLCR IDIQ Agreement 17X074. NCBO is supported by the NIH Common Fund under grant U54HG004028. The CEDAR Workbench is available at https://cedar.metadatacenter.org and on GitHub at https://github.com/metadatacenter.

References


A principled framework for phenotyping postpartum hemorrhage across multiple levels of severity

Matthew Oberhardt, PhD,1, Alexander M Friedman, MD2, Rimma Perotte PhD,1,3, Jean-Ju Sheen, MD2, Alan Kessler, MD4, David K. Vawdrey PhD1,3, Robert Green, MA5,
Mary E. D’Alton, MD2, Dena Goffman, MD2,5

1Value Institute, NewYork-Presbyterian Hospital, New York, NY; 2Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY; 3Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 4Department of Obstetrics and Gynecology, Weill Cornell Medical Center, New York, NY; 5Department of Quality and Patient Safety, NewYork-Presbyterian Hospital, New York, NY

Abstract

Maternal morbidity and mortality have gained major attention recently, spurred on by rising domestic rates even as maternal mortality decreases in Europe. A major driver of morbidity and mortality among delivering women is postpartum hemorrhage (PPH). PPH is currently phenotyped using the subjective measure of ‘Estimated blood loss’ (EBL), which has been shown to be unreliable for tracking quality. Here we present a framework for phenotyping PPH into multiple severity levels, using a combination of data-driven techniques and expert-derived clinical indicators. We validate the framework by predicting large drops in hematocrit and quantitative blood loss, finding that the framework performs better in predicting coded PPH than a hematocrit-based predictor or predictors based on other metrics such as blood transfusions, and does better in predicting quantitative blood loss, a gold standard metric for blood loss that we have for a subset of patients, than any predictor we could build using hematocrit drops alone. In all, we present a principled framework that can be used to phenotype PPH in hospitals using readily available EHR data, and that will perform with more granularity and accuracy than existing methods.

Introduction

Maternal morbidity and mortality have become a topic of great interest recently, with Centers for Disease Control reports of rising US rates (1), higher rates in the US than other developed countries, and recent attention cast on severe maternal morbidity measures by popular media such as the New York Times, Propublica, and USA Today. A major driver of morbidity and mortality among delivering women is postpartum hemorrhage (PPH). Most women bleed after giving birth, but PPH denotes excessive bleeding that can be life threatening if not recognized and treated. As of 2014, hemorrhage is the largest single direct cause of death worldwide among delivering women in both the developing and developed world, with postpartum (as opposed to antepartum or intrapartum) hemorrhage being the most prevalent type of obstetric hemorrhage, the largest individual direct cause of death in the developing world, and the 3rd largest individual direct cause of death in the developed world (2). However despite much effort and attention, PPH rates continue to rise domestically (3,4).

Prevention, improved early recognition, and optimized management are of primary areas of focus for hospitals as they work to improve outcomes for patients, but quality improvement efforts are hindered by the lack of a reliable phenotype for PPH that can be tracked in electronic health records (EHR). The standard definition of PPH is based on ‘estimated blood loss’ (EBL), by which a clinician estimates how many milliliters (mL) of blood a patient has lost after delivery. This metric suffers from basic subjectivity and reliability issues, including a recognized tendency for clinicians to both underestimate and overestimate blood loss in obstetrics (5,6). Nevertheless, EBL is used to determine which patients get the ICD-CM diagnosis code for PPH, and the PPH billing codes are used to track PPH rates for quality analyses, reimbursement, and other needs (6). As with many other diseases, ICD-9 diagnosis codes have been demonstrated to be inaccurate; in this case, ICD codes have been found to underestimate the prevalence of PPH among women undergoing Cesarean delivery (7).
Accurately phenotyping PPH is a first step in understanding causes, prevalence, and potential ways to improve care. Aside from the EBL-based definition of PPH (≥500mL blood after vaginal delivery or ≥1000 mL blood after Cesarean delivery, classically, but modified in 2017 to 1000mL EBL after either type of delivery), PPH has been phenotyped in several ways (6,8). Some guidelines cite Hematocrit (Hct) drops during pregnancy of >10% as a definition of PPH (6,8). However, Hct drops can be affected by many factors including the fluid shifts as a result of intravenous hydration routinely given during and after delivery, hemoconcentration seen in the setting of preeclampsia, as well as blood transfusions given to treat PPH itself. Hct drops are thus fairly non-specific for PPH (we analyze this explicitly in this paper). Other analyses use any transfusions of red blood cells (RBC), or in some cases 4 or more RBC transfusions, as the definition of severe hemorrhage. These definitions are incomplete, as they exclude many deliveries that have large and potentially dangerous PPH, but are dealt with through means other than transfusions (e.g., uterotonics, tamponade), or are otherwise below a severity threshold that would require transfusion. There is another method available for assessing blood loss called ‘quantitative blood loss’ (QBL) which should be considered the gold standard for phenotyping PPH, as it involves objectively measuring the amount of blood lost by a patient. However, this method is fairly new and requires specialized equipment that is not widely adopted in hospitals, and for now, its accuracy can vary depending on the training and approach used for measurement (5,9,10). Until QBL is widely adopted and its application standardized, there remains a need for an electronic phenotyping method that can judge the severity of PPH with available hospital data.

To answer this need, we here present a framework for phenotyping PPH into multiple severity levels, using a combination of data-driven techniques and expert-derived clinical indicators. We believe that a binary classification of PPH (yes/no) is less useful for quality improvement than a more gradated system, which would give a richer interpretation for hospital use, and also would align more with the actual PPH phenotype as captured by QBL: namely, a continuous variable denoting the quantity of blood loss. This suggests that a continuous measure of blood loss would be ideal, but without QBL being routinely measured and with Hct being unreliable (as will be shown), a continuous measure becomes impractical. With 2 severity levels being too few and a continuous variable being impractical (at least until QBL becomes routine), we used clinical judgement to settle on an electronic phenotype for PPH that includes 5 severity levels. The framework is derived based on aggregate Hct drops for patients with various conditions and interventions, the records of which are routinely available in the EHR, plus expert judgement (e.g., for incorporating interventions that affect Hct levels). We also derive an optimal cutoff of Hct drop for phenotyping PPH to compare in performance to our framework. Finally, we validate our method using QBL measurements from our EHR. In all, we present a principled, portable, and data-driven method for phenotyping PPH, which relies on readily available hospital data.

Methods

Cohort
The framework was developed using data from a cohort of all patients delivering babies between July 2014 and July 2017 in four hospitals within a single health system in New York City, numbering 43,657 total delivery visits. Approval to use this data for research was obtained from the Columbia University Institutional Review Board.

EHR data
Data were extracted through automated query from our institution’s electronic data warehouse, which aggregates data from the EHR, billing system, and other sources. Quality checks were performed to ensure that data were not missing, and when necessary, data were pulled directly from the EHR in order to fill gaps. For example, quality checks revealed that uterotonics data were missing in our data warehouse before a certain date from one of our hospitals; we pulled the data directly from the EHR to fill the gaps.

Hct drops
Hct measurements were taken from lab results in the EHR. Hct drops were calculated as the percent decrease between the lowest post-delivery Hct measurement and the last pre-delivery Hct measurement (with a decrease denoting a positive Hct drop).

Definition of the PPH severity framework
A PPH severity framework was built including 5 severity levels. The levels were defined as follows:

- **PPH Level 0**: No PPH
- **PPH Level 1**: Use of uterotonic medications, aside from Pitocin, as documented in the EHR (search terms: methergine, ergonovine, cytotec, misoprostol, hemabate, and carprofen).

- **PPH Level 2**: ICD-9 or ICD-10 diagnosis code for PPH (ICD-9: 66600, 66602, 66604, 66610, 66612, 66614, 66620, 66622, 66624, ICD-10: O720, O721, O722).

- **PPH Level 3**: transfusion of 1-3 units red blood cells as documented in the hospital’s bloodbank, or receipt of a Bakri balloon, which is a uterine tamponade device available across our health system, as documented in EHR notes (search term: Bakri).

- **PPH Level 4**: ≥4 units red blood cells, or hysterectomy on the delivery visit (ICD-9: 68.29, 68.39, 68.49, 68.69; ICD-10: 0UT90ZZ, 0UTC0ZZ, 0UTC70ZZ, 0UBC0ZZ, 0US90ZZ).

Each patient was placed into only one level of the PPH framework on a given visit, corresponding to the highest PPH level they qualify for. So, for example, a patient who received both uterotonics and a Bakri balloon would be placed in PPH level 3 only.

**Logistic regression analysis**

To compare performance of our PPH severity phenotyping model against other methods in predicting a large Hct drop, we built a logistic regression model that predicted ≥28.6% Hct drop (the ‘silver standard’ developed in Figure 1) based on 6 variables: uterotonic administration (aside from oxytocin), coded PPH, Bakri balloon, any transfusion, number of transfusions, and hysterectomy.

**Quantitative blood loss**

QBL readings were extracted from flow sheet data in the EHR. Readings were only available in 2 of our 4 hospitals, and only for part of the time range of the study (January 2016 – July 2017), as the technology is not yet broadly implemented in our hospital system. In quality checks, we noticed that a high percentage (44%) of all QBL readings were divisible by 50. We excluded these readings, assuming that they were actually EBLs that had been inadvertently entered into the QBL flow sheet. After this filtering, there remained 285 QBL values for distinct visits.

**Data handling & analysis**

All data cleaning was done in SQL and R, and data analysis was done in R.

**Framework validation against QBL**

As a validation of our framework, we compared its performance in predicting QBL vs. the performance of analogous, 5-level severity frameworks that are based solely on Hct % drops. We focused on the 255 delivery visits for which we had both QBL and Hct measured (excluding one outlier measurement).

Patients were split into five severity levels so that the Hct-based measure could be fully analogous to our severity framework (which also includes 5 PPH levels when counting “no PPH”). To construct an Hct-based framework, we chose 4 random Hct cutoffs to categorize patients into 5 severity groups, restricting each group to contain at least 10 patients. We did this 1000 times with random Hct cutoffs, so that we could understand how Hct-based 5-level frameworks perform versus our PPH severity framework. To assess ‘goodness’ of a given Hct-based framework, we calculated it’s R^2 value in predicting QBL levels. We present the distribution of R^2 values from different Hct-based frameworks, versus the R^2 value from the PPH severity framework that we build based on clinical records of diagnoses and interventions.

**Results**

**How best to identify PPH using Hct drops?**

Our point of departure in this study is Hct drop, because Hct is ubiquitously measured and could be used universally as a phenotype for PPH if it were sufficiently accurate. Hct drop has been cited before as a proxy for PPH, typically with a cutoff of 10%. A natural question to ask with any new phenotyping method for PPH is: how does it compare to using Hct?

To answer this, we tuned a cutoff of Hct % drops across our whole cohort of patients such that the F-measure is maximized for predicting coded PPH. F-measure balances positive predictive value and sensitivity to achieve a harmonic mean (Figure 1a). The optimal Hct cutoff for predicting coded PPH, which we consider to be silver standard,
is 28.6 percent Hct drop (Figure 1b-c). We present test statistics for this optimal cutoff, plus several other Hct cutoffs that present whole numbers, one of which (30%) performs similarly.

**a.**

\[
F\text{-measure} = 2 \times \frac{PPV \times Sensitivity}{PPV + Sensitivity}
\]

**b.**

![Hematocrit drop](image)

28.6% = Hct drop cutoff that maximizes F-measure for identifying coded PPH
40% = More conservative cutoff

**c.**

![Hematocrit drop](image)

**d.**

<table>
<thead>
<tr>
<th>Hct drop cutoff used to predict coded PPH</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>28.6%</td>
<td>0.39</td>
<td>0.95</td>
<td>0.25</td>
<td>0.97</td>
<td>0.30</td>
</tr>
<tr>
<td>30.0%</td>
<td>0.34</td>
<td>0.96</td>
<td>0.26</td>
<td>0.97</td>
<td>0.30</td>
</tr>
<tr>
<td>40.0%</td>
<td>0.14</td>
<td>0.99</td>
<td>0.42</td>
<td>0.96</td>
<td>0.20</td>
</tr>
</tbody>
</table>

**Figure 1. Development of an optimal Hct cutoff to predict coded PPH.** (a) Definition of F-measure, the harmonic mean of positive predictive value (PPV) and sensitivity. (b) Hct drops across our cohort are shown for patients with vs. without coded PPH. (c) For different Hct cutoffs, test statistics are shown for predicting coded PPH. The blue cutoff maximizes F-score. (d) Test statistics are listed for 3 Hct cutoffs (NPV = negative predictive value), with arrow colors corresponding to the colors of vertical lines in panels b and c.

**Construction of a multi-level PPH framework**

To go beyond Hct drop in predicting PPH, we developed a framework to phenotype patients into different PPH severity levels. Hct drops were used in the construction of the framework to help rank the severity of patient populations receiving different interventions (Figure 2a). The reasoning behind using Hct levels in this way is that, although they are noisy on an individual patient level, the noise would cancel out over a large population and be informative on aggregate about which interventions were used on more severe patients. However, Hct drops have a fundamental limit once patients become transfused, since RBC transfusions affect Hct levels. We therefore used clinical knowledge to re-rank some of the interventions based on known severity (See ‘Regime 2’ in Figure 2a). The final PPH severity phenotyping framework has 4 levels beyond a base level of no PPH (so 5 levels total; Figure 2b). A patient was only placed into one severity level on a given visit, corresponding to the highest level that they qualify for.
Figure 2. Building a severity framework to go beyond Hct in phenotyping PPH. (a) Interventions and diagnosis of coded PPH are ordered according to median Hct % drop among patients in our cohort. Two ‘regimes’ are defined; the first denotes interventions that should be ordered by Hct drop, and the second denotes interventions for which Hct is unreliable, and thus clinical judgement is used to re-rank severity. (b) The resulting PPH severity framework, with Hct distributions shown.

Performance of the PPH severity framework vs Hct
With the PPH framework and optimal Hct cutoff in place, we are positioned to assess the performance of the framework in identifying PPH. For this, we compared several methods for predicting patients with PPH according to the criterion of having Hct drop ≥28.6%, as this was determined earlier to be the most informative Hct cutoff for denoting PPH. We built a logistic regression model using all factors composing the PPH framework (see Methods), and then compared its performance in predicting patients above the Hct cutoff versus several other predictors that were subsets of the framework (Figure 3a). We found that the full framework was more predictive than any subset component, including coded PPH, which notably was the factor that the Hct cutoff of 28.6% was tuned on. To compare the logistic regression results directly against binary factors (such as coded PPH), we report in Figure 3a the logistic regression cutoff that would maximize the F-measure; the full ROC curve for this model is shown in Figure 3b, with modest AUC of 0.68.

Factor used to predict >28.6% Hct drop
PPH framework constituents* 0.36 0.97 0.44 0.95 0.39
PPH level ≥ 3 0.24 0.99 0.64 0.95 0.35
Any units RBC transfused 0.22 0.99 0.68 0.95 0.34
PPH, ICD coded 0.25 0.97 0.39 0.95 0.30

Figure 3. Performance of PPH framework in predicting high Hct drops. (a) We compare different metrics, all of which are included in the PPH phenotyping framework, for their ability to predict Hct drop of >28.6% (see Figure 1). Test statistics are shown. ‘PPH framework constituents’ (denoted with a *) is a logistic regression model with cutoff tuned to maximize F-score in predicting the Hct drop cutoff. (b) ROC curve for the asterisked logistic regression.

Validation of framework vs. QBL
The gold standard for measuring PPH is QBL. We do not routinely measure QBL in our hospital system, although a small subset of deliveries in our cohort (0.6% of total) had QBL recorded (see Methods). We therefore re-examined our framework in light of these QBL measurements, and found that ranking interventions by median QBL recapitulates the exact ordering of interventions that occupy the levels of our PPH framework (Figure 4a). Importantly, the QBL
does not require a ‘reshuffling’ of some PPH interventions to form severity levels based on clinical knowledge (as Hct did: see Figure 2a), but rather it ranks interventions exactly in accordance with the framework. Unlike Hct drops (which cease to be informative once blood is transfused and thus show similar Hct distributions for PPH level 3, which includes <4 units RBC, and PPH level 4, which includes ≥4 units RBC), QBL distributions show a clear separation between PPH levels 3 and 4 (Figure 4b).

![Figure 4. Validation of PPH framework vs. QBL.](image)

*denominators are # visits from the 2 Hospitals from which we could query QBL.

PPH severity framework vs. an Hct-based framework with 5 levels

A final validation of our framework comes in the form of comparing its performance to Hct % drop cutoffs in a fully analogous framework. Since our severity framework contains 5 PPH levels (no PPH, PPH level 1, level 2, level 3, and level 4), we constructed predictors of QBL using Hct cutoffs that split the 285 deliveries for which we had QBL into 5 severity levels, using randomly chosen Hct cutoffs (methods). Fit of a framework was judged by its $R^2$ value in predicting QBL. The best of the 1000 Hct-based predictors is shown in Figure 5a, with an $R^2$ of 0.27. The PPH severity framework we present has an $R^2$ to predict QBL of 0.38, far surpassing any Hct-based predictor that could be generated (Figure 5b). A linear fit of Hct against QBL achieved $R^2$=0.21. These results reinforce that Hct is a noisy signal for blood loss, and the extra information embedded in our PPH severity framework is meaningful enough to boost it considerably beyond what Hct can do.
Figure 5. Performance of PPH Severity framework vs. a Hct-based framework  
(a) The set of cutoffs in Hct to form a 5-level system (i.e., to be comparable with ours: PPH0 – PPH4) that perform the best in predicting QBL.  
(b) The distribution of R^2 values from plots like on the left (from 1000 random Hct cutoff thresholds), versus the R^2 of the PPH severity framework.

Discussion

Importance of measuring

Many efforts are currently aimed at improving the quality of care for women with obstetric hemorrhage, as clinicians, hospitals, states, and national bodies work to decrease severe maternal morbidity and mortality. To understand the impact of these interventions, it is critical that organizations have reliable, easily accessible data that goes beyond using billing codes to demonstrate that a woman bled, required a blood transfusion, or had a hysterectomy. Proposed metrics such as ICU admission and massive blood transfusion are important but also rare, making them less useful as the only measures. The approach described herein provides a practical option for obtaining important clinical data that allows evaluation of process and outcome metrics for the treatment of obstetric hemorrhage. The framework allows assessment of early stage therapies and escalating clinical interventions targeted at stopping bleeding. Clinically, the hope is that as early aggressive interventions are increasingly employed, later stage morbidities associated with hemorrhage will be seen less frequently.

Strengths and weaknesses of framework

The PPH severity phenotyping framework has the weakness that it relies on outcome and process metrics as heuristics to judge whether a patient lost blood. Given, for example, rising PPH rates in a hospital, this makes it difficult to determine whether patients are bleeding more over time, versus more interventions being done on similar patients. Ways to better understand this might include assessing the underlying Hct drops for patients at a given intervention level (although Hct is especially unreliable after transfusion), or assessing potential under-treatment of PPH by looking, e.g., for cases of low fibrinogen levels as a reflection of patients developing coagulopathy in the setting of treatment delays. The most reliable way to move past these challenges will likely be broadly adopting QBL measurement.

In the meantime, the best practices for identifying PPH are woefully inadequate, with subjective measures (EBL) the dominant method. In this landscape, despite its weaknesses, our framework stands as the most principled method to date for phenotyping and analyzing severity of PPH. The framework also has the potential to utilize the EHR to reliably assess and track many process metrics that will be useful to teams seeking to identify whether best practices for treating postpartum hemorrhage are being adopted in their hospitals. Better assessment of process metrics should lead to an impact on outcome metrics for our patients.

Future Work

Planned future work based on the framework developed here includes multiple important efforts. First, we will use this framework to look at trends over time across our healthcare system. Active efforts have been underway to standardize and optimize PPH management through implementation of national and statewide bundles. Utilization of the principled framework will allow assessment of our bundle implementation by evaluating process and outcome measures for women with obstetrical hemorrhage. Only once we have this sort of data will we be able to proactively identify additional improvement opportunities and develop targeted strategies. Secondly, we intend to utilize the framework to look for any variation in PPH practice patterns across our system. Understanding how successfully we have standardized our management will similarly be invaluable to guide further system wide efforts. Finally, there are efforts underway to develop an obstetric hemorrhage dashboard that can be used to follow postpartum hemorrhage management and outcomes in an ongoing fashion. A dashboard including this level of granularity, rather than ICD coded hemorrhage, will be useful to observe the impact of ongoing PPH improvement efforts over time, as well as to ensure sustainability of the processes put into place to minimize morbidity and enhance patient outcomes.

Conclusion

The framework developed and presented in this work is a reflection of the power of collaboration between engaged clinicians, quality improvement team members, and data-analytics experts. None of these groups in isolation has the power to develop such a tool without input from the rest of the team. Individuals with the expertise to obtain this sort of data from the EHR can help drive clinically important improvement work with the potential for large-scale impact.
on patient outcomes. The presented phenotyping framework, a culmination of this tight collaboration, should provide a strong tool for future use in tracking obstetric hemorrhage.

Acknowledgements
Dr. Friedman is supported by a career development award (K08HD082287) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health.

References
How Do Spinal Surgeons Perceive The Impact of Factors Used in Post-Surgical Complication Risk Scores?

Enea Parimbelli, PhD¹, Szymon Wilk, PhD¹,², Dymphna O’Sullivan, PhD³, Stephen Kingwell, MD⁴, Wojtek Michalowski, PhD¹, Martin Michalowski, PhD⁵
¹University of Ottawa, Ottawa, ON, Canada; ²Poznan University of Technology, Poznan, Poland; ³Technological University Dublin, Dublin, Ireland; ⁴The Ottawa Hospital, Ottawa, ON, Canada; ⁵University of Minnesota, Minneapolis, MN, USA

Abstract

When deciding about surgical treatment options, an important aspect of the decision-making process is the potential risk of complications. A risk assessment performed by a spinal surgeon is based on their knowledge of the best available evidence and on their own clinical experience. The objective of this work is to demonstrate the differences in the way spine surgeons perceive the importance of attributes used to calculate risk of post-operative and quantify the differences by building individual formal models of risk perceptions. We employ a preference-learning method - ROR-UTADIS - to build surgeon-specific additive value functions for risk of complications. Comparing these functions enables the identification and discussion of differences among personal perceptions of risk factors. Our results show there exist differences in surgeons’ perceived factors including primary diagnosis, type of surgery, patient’s age, body mass index, or presence of comorbidities.

Introduction

When deciding about surgical treatment, an important factor driving a surgeon’s decision-making is the perceived risk for post-surgical complications. Spine surgery is no exception, especially considering that introduction of new surgical implants makes even more patients eligible for spine surgery¹. Moreover, significant variability is observed in the use of spine surgery, among different hospitals, as well as at the level of an individual surgeon²–³. Because risk is inherent in any procedure, reducing the number of unnecessary, or high-risk operations is an important issue in patient safety and will improve overall patient outcomes, reduce complication rates, and reduce the need for repeated surgery.⁴ Thus, the ability to correctly and consistently assess the risk of post-surgical complications plays an important role in deciding if a patient is eligible for surgery and what type of surgical procedure should be considered.

A spinal surgeon’s risk assessment is based on their knowledge of the best available evidence as well as on their own clinical experience. Commonly used risk assessment tools in spinal surgery include SpineSage⁵ and spinalRAT⁶, and we use these as baselines in this work. These tools consider patient demographics and additional attributes such as patient condition, comorbidities, pre-surgical diagnoses, surgical and complication detail, and others. Although little is known about how surgeons weigh available evidence and clinical experience, emphasis on experience is likely increased when little scientific evidence exists, the surgeon is unaware of that evidence, or the surgeon has discounted that evidence based on its quality. For these reasons, differences exist among spinal surgeons while assessing the risk of post-surgical complications for the same patient. These differences translate into uneven application of standards of care and inconsistent selection of patients eligible for surgery. Identification and analysis of differences should help mitigate these inconsistencies and is the topic of the research described in this paper.

The objective of our work is to identify how spinal surgeons weigh the importance of the variables in risk assessment tools. Towards this end we assess the differences in the way spinal surgeons assess risk by building formal models of risk perceptions of 6 spinal surgeons working in a large academic hospital in Canada. We employ a preference-learning method - ROR-UTADIS⁷- to build preference models that capture surgeon-specific perception of risk of complications. We previously used a similar approach to capture the professional opinions of physicians when evaluating the relevance of medical evidence for decision making⁸. While other research on preference elicitation involving surgeons has focused on treatment options⁹–¹², our work is more concerned with identifying the differences in post-surgical risk assessment and attempts to discover the most relevant clinical factors, and how surgeons balance them when making such an assessment.

Related Work

Surgical decision-making has evolved over time. What was once an intuitive matter for surgeons has now become a complex multi-faceted decision process¹³. Risk assessment has a substantial subjective element, thus there is a need
for objective tools and methods. Such approaches aid surgeon decision making but also provide realistic expectations for the patient, helping them make an informed decision.

Scoring systems that focus on postoperative outcomes are a common method for predicting risk. For example, www.riskprediction.org.uk, lists 12 risk assessment tools related to various procedures. Scoring tools are generally based on prognostic factors including age, disease severity and co-morbidity and procedure specific considerations are available to surgeons. For example, the well-known POSSUM system (Physiological and Severity Score for the Enumeration of Mortality and Morbidity), and its variants, compute information on the surgical risk in terms of morbidity and mortality by combining physiological parameters (e.g. age, cardiac status, respiratory status) with operative parameters (e.g. the type of operation, the urgency of the operation and the number of procedures).

Spinal surgery is a complex procedure and requires clinicians to evaluate the relative risk of several risk factors including age, gender, medical comorbidities, substance abuse, body mass index (BMI), medical comorbidity, previous spinal surgery, primary diagnosis and surgical approach. Given the large number of possible risk factors, many approaches to assessing spinal surgery risk have used statistical techniques such as regression and multivariate analysis. Although predictions from systems such as those described by can be individualized, they largely pertain to populations rather than an individual. SpineSage is a tool that attempts to individuateize the risk assessment of spine surgery by taking into consideration patient-specific risk factors, specifically a patients’ comorbidity profile as well as the invasiveness of the procedure. It utilizes a multivariate log-binomial approach. Other work has employed machine learning, for example Ehlers et al. have developed a Naive Bayes algorithm that uses 300 predictors to predict risk of adverse event or death within 90 days of different types of surgery including spine surgery. Karhade et al., compared four machine learning algorithms for preoperative prediction of non-routine discharges for elective inpatient lumbar degenerative disc disorders and found a neural network approach to be the best performing with an AUC of 0.823.

In addition to the large number of factors to be evaluated, there are other emerging issues that cause poor consistency in risk assessment for spine surgery. These include an increasing number of available surgical interventions with collateral adverse outcomes that may be traded off against each other. Furthermore, risk factors vary considerably with comorbidities and different patient populations present very specific concerns. For example, surgical risk is especially challenging in elderly patients because of their levels of frailty which is not commonly considered by risk prediction systems and scoring systems or machine learning algorithms based on population level characteristics do not generalize well to all patient groups. Our approach in analyzing the preferences of a number of surgeons for multiple and heterogeneous risk factors aims to shed light on what the important decision-making inconsistencies are among spinal surgeons. We believe this is first time such an analysis has been conducted in the spine surgery domain.

Methods

The ROR-UTADIS method

In our work we assume a preference model that captures perceptions of risk of complications and is represented as an additive value function. An additive function is the sum of marginal value functions associated with specific criteria characterizing alternatives. Here alternatives correspond to patients and criteria to variables characterizing patients that are listed in Table 1. The additive value function not only provides a comprehensive assessment of a patient (in terms of perceived risk) but also, through marginal value functions, gives insight into risk perceptions associated with individual variables which is crucial for achieving our research goals.

Most methods for building additive value functions establish parameters of marginal value functions from indirect preferential information provided by a decision maker for a subset of alternatives (so-called reference alternatives), e.g., their pairwise comparisons. Usually, there are multiple additive value functions compatible with preferential information. A simple approach involves selecting any of these functions, while an advanced one takes into account and exploits all these compatible functions for more robust results. The latter approach is known as robust ordinal regression (ROR).

In our study we employed the ROR-UTADIS method that follows the ROR principle. It aims to solve a sorting decision problem, i.e., assignment of alternatives to predefined and ordered classes (e.g., risk classes). In addition to a resulting additive value function it also establishes a set of thresholds that can be imposed on obtained quantitative assessments to translate them into class recommendations. ROR-UTADIS accepts rich and diversified preferential information, including class assignments (possibly imprecise) of references alternatives, their assignment-based pairwise comparisons, and desired class cardinalities. In our analysis we focused solely on the value function and...
ignored the thresholds. Moreover, we used only the first type of preferential information, i.e., assignment of reference (paper) patients to risk classes, as it was most relevant from a practical perspective and easiest to obtain from surgeons.

Below we briefly summarize the outline of the ROR-UTADIS method. In this description we use terminology specific for our clinical problem:

1. A set of value functions that are compatible with provided assignments of patient cases to risk classes (low, medium and high) is constructed by solving a linear programming model. If this model has no solution, it indicates that some of the provided class assignments are problematic and need to be revised, and the method proceeds to step 2. Otherwise, it skips to step 3.

2. Problematic class assignments are identified by solving a 0-1 linear programming model. Found class assignments are presented to a decision maker who needs to decrease their precision, e.g., an initial assignment of a specific patient to the medium risk class is changed into an assignment to medium or high classes (the obtained solution also points to whether the assignment should be expanded towards a lower or higher risk). Once all problematic assignments have been revised, ROR-UTADIS returns to step 1.

3. A set of compatible additive value functions established in step 1 is explored and a representative additive function is constructed. This step involves solving another linear programming model that is aimed at maximizing the differences in evaluations between patient cases assigned to different risk classes and then minimizing differences in evaluations between patients assigned to the same risk classes.

Study Design

Our study population consists of 6 staff surgeons (4 orthopedic and 2 neurosurgical surgeons) from the Division of Orthopedic Surgery at The Ottawa Hospital (TOH), Ottawa, Ontario, Canada. The participants are academic, fellowship-trained spinal surgeons, with training at Canadian, American, British, and Australian hospitals. All of the surgeons work closely in a combined orthopedic and neurosurgical spine program. This study population excluded residents and included all but one of the staff spine surgeons at TOH. While all 7 staff surgeons agreed to participate, one did not provide answers. Yet this study population represents almost the entirety of the TOH spine program and thus represents the group’s thoughts. All the surgeons were considered to have the same experience level.

Table 1. List of features considered in the study.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Feature name</th>
<th>Value domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure type and approach</td>
<td>Proc_Approach_CdGrp</td>
<td>1=Anterior cervical 2=Posterior cervical 3=Posterior thoracolumbar</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>MRDx_10_CdGrp</td>
<td>1=Degenerative 2=Trauma</td>
</tr>
<tr>
<td>Diabetes</td>
<td>ncdDiabetes</td>
<td>1=Yes, 0=Otherwise</td>
</tr>
<tr>
<td>Hypertension</td>
<td>ncdHypertension</td>
<td>1=Yes, 0=Otherwise</td>
</tr>
<tr>
<td>Bleeding diathesis</td>
<td>ncdBleeding</td>
<td>1=Yes, 0=Otherwise</td>
</tr>
<tr>
<td>Age</td>
<td>Age</td>
<td>18-100</td>
</tr>
<tr>
<td>BMI</td>
<td>ncdBMI</td>
<td>15-50</td>
</tr>
</tbody>
</table>

Study participants were provided with a set of 15 representative hypothetical patient cases described by a set of 7 features that are available at the time of first consult. First consult is typically when surgery is planned and considered when evaluating risk of post-surgical complications (see Table 1). These features include type of the surgical procedure and how it was conducted (often called “approach”), diagnosis, features associated with comorbidities such as diabetes, hypertension, and known bleeding diathesis, age, and BMI. The patient cases were developed by an independent spinal surgeon on the basis of cases recorded in The Ottawa Hospital Institutional National Surgical Quality Improvement Program (NSQIP) database for patients undergoing spinal surgery. The patient cases were
developed based on EHR data and represent the patient population that is being treated in the academic center. This population covers different age groups, resulting in the age variation seen across patient cases.

As a starting point, we elicited individual risk assessments from surgeons asking them to assess the risk of post-surgical complications for each patient case on a low/medium/high risk scale. In the first stage of the analysis assessments were used to construct a common additive value function representing a perception of risk by all participating surgeons. The second stage of analysis focused on building a set of surgeon-specific additive value functions to highlight differences in individual risk assessment.

**Results and Discussion**

The patient cases considered in this study and the assignments to risk classes provided by the participating surgeons are given in Table 2. As summarized in the bottom portion of the table, there exists a fairly large number of disagreements among surgeons with regards to the perceived risk of post-surgical complications. Indeed, only one patient case (P6) had the same risk assessment from all spinal surgeons, while one third (5/15) of the patients had risk assessments that span the whole spectrum from low to high risk.

**Table 2.** Risk assessment of the 15 paper cases (P1-15 in columns) by the 6 spine surgeons (S1-6 on rows). The bottom part of the table reports frequency of risk classes for each patient when pooling all surgeon assessments together.

In the first stage of the analysis, we combined individual assessments to build an overall additive value function capturing common risk perception by all participating surgeons. Merging responses provided by specific surgeons resulted in imprecise risk class assignments that were provided as input to the ROR-UTADIS method. For example, patient P1 was assigned to low or medium risk classes, P2 to all risk classes, and P3 to medium and high risk classes. Figure 1a shows the marginal risk functions obtained for the combined assessments. Each marginal function describes how values of a specific feature contribute to the perception of risk with feature values being reported on the X-axis and the contribution on the Y-axis. The latter is expressed using a scale from 0 (no contribution) to 1 (maximum risk).
Figure 1. Marginal risk functions for combined assessments of all surgeons (a) and for individual surgeons (b). Functions on a single row refer to a specific feature, while columns represent either the group of all participating surgeons or individual surgeons (S1-S6).

Findings from this stage of our study align with common clinical practice and general guidelines for spine surgery. Procedure type/approach and diagnosis all emerged as contributing to the risk of post-surgical complications. The
same can be said about diabetes. Age and BMI contributions depend on values recorded for a patient. Initially for low values these two features are not contributing to risk, this starts to change after a threshold of 60 years of age and BMI of 25 (this is aligned with what clinical guidelines state), to finally become major drivers for risk of complications for high values - age greater than 85 and BMI greater than 30. Hypertension and bleeding diathesis do not impact the perceived risk of post-surgical complications.

In the second stage of the analysis we applied ROR-UTADIS again in order to discover if individual surgeons differed in the ways they evaluate the importance of specific features and their values. The method was not initially able to construct additive value functions for S3, S4, S5 and S6 and identified problematic class assignments for some patient cases (P2, P4, P7, P8, P9, P11 and P14). These assignments had to be revised by making them less precise (e.g., the risk class assignment of P2 by S3 was changed from medium to low or medium), while the remaining precise assignments reported in Table 2 were kept unchanged.

Surgeon-specific marginal value functions are reported in Figure 1b. Interestingly, surgeon-specific models highlight how the two categorical features characterizing the type of procedure type/approach and the diagnosis have different impact on risk according to individual surgeons. S6 does not attribute any additional risk to the different procedure type/approach categories used in this study; S1 and S2 consider anterior cervical surgeries to be associated with smaller risk and attribute the same risk to the surgeries with posterior approach; finally, S3, S4 and S5 consider thoracolumbar surgeries considerably more risky than cervical ones. A somewhat similar consideration is true for the diagnosis feature where S4 and S5 consider operating on patients with degenerative and traumatic lesions as associated with the same risk, while the other surgeons attribute more risk (with different intensities) to traumatic lesions. These observations are explained by the fact that procedure type/approach and diagnosis together describe the overall complexity of a patient case. However clinical experience of each surgeon and the frequency of operating on a specific group of patients (e.g. trauma) or performing a specific type of surgery (e.g. posterior approaches, or thoracolumbar) are likely to influence the risk propensity that they ascribe to individual features.

Diabetes, hypertension, and known bleeding diathesis are all well-known risk factors but the specific weighting of each of them in the risk assessment process is subjective to the individual surgeon considering both evidence and experience. This is reflected in the fact that only a subset of surgeons attribute marginal risk when any of these features is present. For example, S2, S3, S4 and S5 do not consider operating on diabetic patients riskier, while S1 and S6 attribute an important (0.30) risk contribution to having diabetes. The observed variability of weights attributed to the same feature by different surgeons (e.g. hypertension has a non-zero weight only for S6) is also an example of subjectivity as a result of surgeons’ experience and available evidence.

Features such as age and BMI are also commonly considered risk factors, but little is known as to whether the risk increases linearly, exponentially, or in a stepwise fashion at certain threshold levels. Furthermore, chronological age does not necessarily accurately reflect physiological age or general health. These considerations make the contribution to risk of factors like age and BMI very subjective in terms of cut-off points. Surgeon-specific models shown in Figure 2 reflect this variability, but still highlight the fact that age and BMI influence risk for all but one surgeon. After discussing the results with the participating surgeons, it was confirmed that for some a threshold impacting risk assessment can be set at BMI=25 while for others at BMI=30 or even more. Similarly, some surgeons will not operate on patients over 90 years old as they consider this age group to be of too high risk of post-surgical complications.

While the list of features considered in our study is small compared to existing tools such as SpineSage and spinalRAT, we can still make several observations applicable to these tools. SpineSage uses age, surgical invasiveness score, bleeding disorder, congestive heart failure, and diagnosis of spine trauma and spine infection to predict any complication and age, surgical invasiveness score, gender, chronic pulmonary disease, hypertension, previous cardiac history, and diagnosis of spine trauma and spine infection to predict major complications. spinalRAT uses age, gender, comorbidities, preop diagnosis, location of surgery, use of BMP, fusion status, and instrumentation status for predication. While these features are more comprehensive than ours, we have commonalities (age, comorbidities, diagnosis) in our feature set. These common features, and proxies for others, allow us to deduce that surgeons will perceive the importance of the features in a similar manner when using SpineSage, spinalRAT, and other similar tools.

**Conclusion**

Spinal surgeons’ decision-making process is driven by the scores reported by risk assessment tools they use to quantify post-surgical complications. However, their interpretation and weighing of the importance placed on each attribute used to calculate these scores varies. The relative contribution of each attribute is vital in making accurate predictions.
of complications and should be considered when using these assessment tools for surgical decision-making. As such, the relative importance of the attributes that make up each score, as determined by a surgeon, should be better explained to inform the construction of the surgical plan and to educate the patient.

In this work, starting from data collected from 6 experienced spinal surgeons, we built a formal additive model summarizing contribution of specific factors to the risk of complications for spine surgical patients. The overall model elicited for all surgeons aligns with available scientific evidence, highlighting variables that the majority of surgeons deemed important for risk assessment. However, significant differences were observed in the way individual surgeons assign importance to attributes used to evaluate risk. Our preference-learning-based method proved effective in discovering these differences and enabled in-depth discussion with the clinical experts involved. Results suggest that factors such as overall complexity of the surgery, presence of comorbidities, age, and BMI all play an important but highly subjective role in complication risk assessment for spine surgery and ultimately need to be carefully considered during spine surgery planning.

Our study includes several limitations that are mentioned below:

- The preference model represented as an additive value function assumes independence of different variables and does not consider interactions among them;
- The number of surgeons participating in the study is limited to 6 and all of them practice in the same institution;
- Evaluations were conducted on patient cases that, albeit realistic, constitute a simplified approximation of a real patient case;
- Patient cases evaluated by surgeons excluded day surgeries, which represents a relevant proportion of spine surgical interventions, but are intrinsically low risk.

Future research will address some of the limitations of our current approach. For example, considering rule-based preference models that are able to capture interactions between features, using richer description of patients and involving a larger number of participants. Collecting more preference-oriented data (e.g. including surgeons still in training, as well as across different sites) will result in creating a decision support tool that allows insight into the trade-offs between the different factors involved when evaluating risk.

References

Technology Implementation and Associated Pharmacy Interruptions
Kalyan S. Pasupathy, Ph.D., Mayo Clinic, Rochester, MN
Linsey M. Steege, Ph.D., University of Wisconsin-Madison, Madison, WI
Chris C. Cho, M.S., University of Wisconsin-Milwaukee, Milwaukee, WI

Abstract
This study focuses on interruptions in an inpatient pharmacy setting and the impact of CPOE implementation on the types, frequency, and duration of interruptions. A cross-sectional observation study of pharmacy employees in an inpatient pharmacy was conducted. The independent variables included day of week, time of day, job position of the person interrupted, and description of each interruption. A total of 552 interruptions were observed with a mean frequency of 10.6 interruptions per hour lasting a mean (SD) duration of 1.34 (1.43) minutes. Incoming calls were the most frequent interruption type across all phases. Pharmacy employees spend almost a quarter of their time on interruptions, and pharmacists have longer interruptions than technicians. Immediately after CPOE implementation, durations tend to be one-and-a-half times longer than before. CPOE implementation did not affect the frequency of interruptions. Recommendations included redesign of work processes and job responsibilities.

Introduction
Medication error is a major cause for concern within the healthcare sector have major safety and cost implications. Such errors lead to approximately 9,000 preventable deaths per year and accrue costs due to increased lengths of stay. Hospital pharmacy accounts for approximately 15 percent of a hospital budget. Hence, understanding the root causes of medication errors and changing work processes is vital to improving safety, reducing costs and achieving higher levels of overall quality of healthcare. One type of medication-related error is a dispensing error occurring within pharmacy. Medication dispensing errors can be attributed to several causes such as staffing constraints, excessive workload on pharmacists and technicians, interruptions and distractions, lighting levels, and noise. Interruptions in physicians have been studied and have been associated with errors. Pharmacy interruptions can also impact performance, and such interruptions are a major concern especially with pharmacists and technicians working on visual tasks. Research studies indicate that interruption of an activity may have a negative impact on human performance. This is because the mean processing time is higher for a task that was interrupted than if the task was not interrupted. Further, previous work has found that physicians and nurses generate twice as many interruptions as they receive, leading to an interruptive workplace and contributing to inefficiency in work practice. Since the focus is on continuous quality improvement and value of the care provided to patients, it is important to understand interruptions and their factors, to redesign work processes and improve quality.

Hospitals are implementing new technology such as electronic medical records, Computerized Prescriber Order Entry (CPOE), electronic medication order processing, electronic medication administration record (eMAR), etc. Some authors have proposed that health information technologies ultimately increase clinicians’ access to data and thus may decrease interruptions between clinicians asking for clarification or information. However, others have observed that newly implemented health information technologies, such as electronic medical records or medication administration tools, increase complication of work tasks, disrupt work flow (including increases in interruptions), and do not support clinician information needs. They may also disrupt communication and coordination processes between various clinical personnel. Further, much of the existing work on the effect of implementing health information technologies on work flow has focused on single clinician types or a single unit or work environment. Less is known about the implications to work processes across disciplines and work areas. CPOE, in particular, has been shown to impact work flow and processes beyond physician providers. It is also important to note that computer-based health information technologies are generally not used in isolation; clinicians may regularly be interrupted while performing data entry and retrieval tasks and these interruptions have been cited as a source of increased errors following implementation of health information technology. Hence, it is important to study interruptions with respect to technology implementation beyond isolated professions/disciplines and understand trends and patterns to redesign work processes and improve quality of care.

The overall objective was to study the impact of technology implementation, specifically CPOE, on interruptions in an inpatient pharmacy. The specific research question was to study how technology implementation affects, types of interruptions, frequency of interruptions, and duration of interruptions.
Method
This study used a cross-sectional observation method to record interruption types, times, durations, and pharmacy personnel involved during three data collection periods: prior to CPOE implementation, immediately post CPOE implementation, and approximately 5 months post CPOE implementation. Since the primary focus was to study the impact of CPOE implementation, the sessions were sampled to account for pre-, immediate post and later post implementation, so that the study could capture both immediate and long term effects of CPOE implementation.

Observation sessions: Prior to data collection, the study was approved by the Institutional Review Board (IRB). All pharmacists and technicians agreed to the observers being present. Three observers spent a total of 29 sessions spanning 52.25 hours in the inpatient pharmacy of a 274-bed academic hospital. Each session had one or two observers present. The observation sessions were scheduled for up to two hours each. Sessions were included from each of the three types of day classifications (Mon-Thu, Fri, and Sat-Sun), and three different time of day classifications - morning (8 a.m. to 11 a.m.), afternoon (11 a.m. to 5 p.m.) and evening (5 p.m. to 9 p.m.). These type of day and time of day classifications were identified based on significant differences in medication orders received by pharmacy across these day and time categories. Our assumption here is that based on significant differences in medication orders received by pharmacy, interruptions would be similar within these categories and different between these categories. During each session, the observers situated themselves in pre-determined locations close to the pharmacist/technician and recorded time and duration using stopwatches, and description of the interruptions. The observers focused on behavioral cues to identify interruptions (ceasing a task to respond to external stimuli). When an interruption occurred, the observers used contextual cues to determine the root cause and classified the interruption. The duration was noted until the point in time when the employee either returned to the original task or was further distracted. For cases where the observer was unable to determine the cause of the interruption, the pharmacist or technician was prompted to provide a brief description.

Calibration and inter-observer reliability: Prior to any data collection, two observation sessions were conducted to calibrate the observation process and ensure consistency in the data collection process. During these calibration sessions, all three observers were present and recorded the same set of interruptions. After completion of the second observation session, 100% consistency in the data collection and coding was achieved. To ensure reliability, the data were tested. After accounting for when the sessions were scheduled (day of week and time of day), no significant differences among the observers were identified, thus ensuring inter-coder reliability.

Data Collection: Data from the observation sessions included CPOE phase, type of day, time of day, type of interruption, job title of the person interrupted (job position), duration and description of each interruption. The type of interruption included one of six types – incoming phone calls, outgoing phone calls, discussion, window, door and hands-on task. Incoming phone call interruptions were to clarify medications, locate missing medications, confirm that the orders placed have been received, check-up on the status of the orders placed and hurry-up the processing of the orders placed. Outgoing phone calls were primarily to change a medication due to an adverse drug interaction, allergy conflict, wrong dosage, etc. While other interruptions can be detrimental to patient safety, outgoing phone calls can be argued to improve patient safety. This study does not distinguish the type of interruption as increasing/decreasing risk of errors. Rather, the primary focus is on interruptions that disrupt the smooth flow of operations. Discussion included interruptions when a pharmacy employee had to consult another employee, for instance, a pharmacist instructing the technician to fix a computer or a technician/pharmacist consulting a pharmacist on a medication-related question. Window and door refer to interruptions at the window and door respectively. And finally, hands-on task is an interruption that required the employee to have to make a major move from their location and use both hands for another task like checking the incoming tube, using the print/fax machine, writing on the board, etc. Job position refers to the person interrupted, pharmacist or technician. The description column recorded the reasons for each interruption. This data was recorded in a spreadsheet for analyses. All data variables were coded as nominal, except for duration which was coded as a scale variable. The types of interruptions, frequency and duration were the dependent variables and the rest were independent variables. For each CPOE phase, the difference in the type, frequency and duration of the interruptions were analyzed. Further for each CPOE phase, the difference in the type, frequency and duration of the interruptions by type of day, time of day and job position were also analyzed.

Analysis: The three outcomes of interest were types, frequency and duration of interruptions. Continuous variables are described using mean and standard deviation. And, univariate differences were compared using t-test and Wilcoxon rank-sum test. Nominal categorical variables were described using frequency and percentages and
univariate difference were examined using Pearson chi-square test. Specifically, for the types of interruptions, Pearson Chi-squared test was used for both the unstratified and stratified analyses. For the frequency of interruptions, there were 29 sessions with varying session duration lengths. Therefore, for this analysis, a Poisson regression model was used where the session duration with an offset was accounted for via an offset parameter and the over-dispersed variance was scaled using the deviance. We evaluated the frequency of interruption for a given session and the factors included day of the week (Mon-Thu, Fri, Sat-Sun), time of day (morning, afternoon, evening), CPOE Phase (Pre-CPOE, Immediate Post-CPOE, Post-CPOE). In addition, the difference in the variation of the rate of interruption was assessed using the Levene’s test for homogeneity.

For the duration of the individual interruptions, which was measured in number of minutes, the data was transformed via a natural log function because the original form of the data was not normally distributed. A multiple linear regression model was used to analyze the log-transformed duration of the interruptions and the variables examined include day of the week, time of day, type of interruption, job position, CPOE Phase and all pairwise interactions. For both the Poisson regression and linear regression models, a backwards elimination method with a p-value criterion of 0.05 was used for model selection.

Avoidance of bias: The sessions in the three phases were of different total lengths, however this will not affect the percent of types of interruptions observed and the duration of interruptions observed. The different lengths may affect the frequency of interruptions observed, which was addressed by normalizing the number of interruptions in any given session using the length of the session to compute the frequency of interruptions per hour.

Results
A total of 552 interruptions were observed at a mean frequency of 10.6 interruptions per hour and lasting a mean (SD) duration of 1.34 (1.43) minutes. This translates to spending 14.2 minutes dealing with interruptions for every hour of work in pharmacy. Or in other words, pharmacists and technicians working in order processing spend almost a quarter (25 percent) of their time on interruptions. Descriptive statistics of the sessions are provided in Error! Reference source not found. below, and those of interruptions are provided in Table 2.

Table 1. Descriptive Statistics on Interruptions by Observation Sessions

<table>
<thead>
<tr>
<th>Phase</th>
<th>Total Time Period of Sessions (min)</th>
<th>No. of Interruptions</th>
<th>Mean Interruptions per hour</th>
<th>Mean (SD) Duration per Interruption (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-CPOE</td>
<td>1230</td>
<td>227</td>
<td>11.1</td>
<td>1.1 (1.16)</td>
</tr>
<tr>
<td>Immediate-Post CPOE</td>
<td>905</td>
<td>148</td>
<td>9.8</td>
<td>1.7 (1.88)</td>
</tr>
<tr>
<td>Later-Post CPOE</td>
<td>1000</td>
<td>177</td>
<td>10.6</td>
<td>1.3 (1.25)</td>
</tr>
<tr>
<td>Type of Day</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mon-Thu</td>
<td>1185</td>
<td>204</td>
<td>10.3</td>
<td>1.4 (1.39)</td>
</tr>
<tr>
<td>Fri</td>
<td>1045</td>
<td>186</td>
<td>10.7</td>
<td>1.2 (1.46)</td>
</tr>
<tr>
<td>Sat-Sun</td>
<td>905</td>
<td>162</td>
<td>10.7</td>
<td>1.3 (1.47)</td>
</tr>
<tr>
<td>Time of Day</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morning (8 a.m. - 11 a.m.)</td>
<td>1035</td>
<td>183</td>
<td>10.6</td>
<td>1.0 (0.95)</td>
</tr>
<tr>
<td>Afternoon (11 a.m. - 5 p.m.)</td>
<td>965</td>
<td>198</td>
<td>12.3</td>
<td>1.3 (1.41)</td>
</tr>
<tr>
<td>Evening (5 p.m. - 9 p.m.)</td>
<td>1135</td>
<td>171</td>
<td>9.0</td>
<td>1.7 (1.77)</td>
</tr>
</tbody>
</table>

Types of interruptions:
When the types of interruptions were analyzed using multiple chi-squared tests, no significant differences were observed across the three phases. However, within the three phases, several significant differences were found based on type of day, time of day and the job position. The type of interruptions was significantly different by type of day only for pre-CPOE and later post-CPOE phases. Interruptions during the immediate post-CPOE phase did not have any significant differences based on the type of day, see Table 3. Incoming calls were the highest across all day
types and phases. Otherwise, changes were noted with outgoing phone call, window and discussion interruption types. Window interruptions moved from being highest on Monday-Thursday pre-CPOE to Saturday-Sunday during later-post CPOE. Outgoing phone calls moved from being highest on Saturday-Sunday and Monday-Thursday during pre-CPOE to Friday during later post-CPOE phase. Discussion interruptions moved from being highest on Saturday-Sunday pre-CPOE to highest on Monday-Thursday during the later post-CPOE phase.

Table 2. Descriptive Statistics on Interruptions by Other Categorizing Variables

<table>
<thead>
<tr>
<th>Type of Interruption</th>
<th>Number of Interruptions</th>
<th>Percent Interruptions</th>
<th>Mean (SD) Duration per Interruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incoming phone calls</td>
<td>380</td>
<td>68.8%</td>
<td>1.3 (1.32)</td>
</tr>
<tr>
<td>Outgoing phone calls</td>
<td>52</td>
<td>9.4%</td>
<td>1.4 (1.21)</td>
</tr>
<tr>
<td>Discussion</td>
<td>67</td>
<td>12.1%</td>
<td>1.3 (1.78)</td>
</tr>
<tr>
<td>Window</td>
<td>33</td>
<td>6.0%</td>
<td>1.9 (2.26)</td>
</tr>
<tr>
<td>Door</td>
<td>11</td>
<td>2.0%</td>
<td>0.6 (0.54)</td>
</tr>
<tr>
<td>Hands-on task</td>
<td>9</td>
<td>1.6%</td>
<td>1.3 (0.97)</td>
</tr>
</tbody>
</table>

Job Position

| Pharmacist                 | 290                     | 52.5%                 | 1.5 (1.45)                          |
| Technician                 | 262                     | 47.5%                 | 1.2 (1.40)                          |

The time of day was significant only during the pre-CPOE phase, see Table 3. Again incoming calls was the highest type of interruption across all times of day. Otherwise, discussion interruptions tend to be higher during mornings and afternoons, while outgoing calls tend to be higher during mornings and nights. The time of day was not significant for immediate post-CPOE and later post-CPOE phases. The types of interruptions received by pharmacists and technicians across all three phases were also significantly different, Table 3. Specifically, regardless of CPOE implementation phase, incoming calls were the highest for both pharmacists and technicians. In addition, when pharmacists were interrupted, they typically tended to have discussion or outgoing phone call interruptions. Whereas, technicians were interrupted more by incoming phone calls and window interruptions.

Frequency of interruptions:
To analyze the frequency of interruptions, the frequency was computed for each session based on the number of interruptions that occurred and the total duration of that session. When frequency of interruptions (number of interruptions per hour) was analyzed no significant relationships were identified. In other words, there is no significant evidence to suggest that frequency of interruptions were different for the three CPOE phases. On average, the pharmacy experiences approximately 10.12 (SD = 3.67) interruptions per hour during the course of its daily operations. The rate of interruptions tends to vary slightly and increases from an average of 9.51 interruptions per hour in during Mondays-Thursday to 10.84/hour in the weekends. However, this difference is not statistically significantly (p=0.7707). In terms of time of day, the opposite trend is observed where on average there are 10.92 interruptions per hour during the morning and decreases to 8.96 per hour in the evening. Similarly, this decrease is also not significant with a p-value of 0.5738. Through the three phases of CPOE, an increasing but not significant rate of interruptions per hour was observed from 9.63 per hour during the pre-CPOE phase, to 10.18 per hour during the immediate post-CPOE phase and to 10.65 per hour at the later post-CPOE phase (p=0.8028). However, significant differences in the variation of the rate of interruption was found across the three CPOE phases (p=0.0384). Specifically, the variation in the rate of interruptions increased significantly from pre-CPOE to immediate post-CPOE to post-CPOE phases. Please note that differences in variation of interruption rate across days of the week and times of the day were not significant.

Duration of interruptions:
When duration of interruptions was analyzed, there were significant differences (p-value=0.0002) between the three phases (pre-, immediate-post and later-post). The mean duration per interruption increases from 1.1 (1.16) minutes during pre-CPOE to the highest level of 1.7 (1.88) minutes during immediate post-CPOE and drops back to 1.3 (1.25) minutes during later post-CPOE. Time of day was a significant factor during the pre and immediate post
phases, with mornings having the shortest duration and evenings with the longest duration of interruptions. Another important factor affecting the duration of interruption was the job position, technician interruptions tend to be shorter than pharmacist interruptions. This was significant across pre-CPOE and later post-CPOE phases. Both pharmacist and technician interruptions increased to the highest level during the immediate post-CPOE phase. While the pharmacist interruption duration increased 58 percent, the technician interruption duration increased only 22 percent. Type of day did not have any significant effect on the duration of interruptions in any of the three CPOE phases.

Table 3. Types of Interruptions by Phase and Type of Day, Time of Day and Job Position

<table>
<thead>
<tr>
<th>Phase</th>
<th>Type of Day</th>
<th>Type of Interruption</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Incoming phone calls</td>
<td>Outgoing phone calls</td>
</tr>
<tr>
<td>Pre-CPOE</td>
<td>Pharmacist</td>
<td>58.1%</td>
<td>17.1%</td>
</tr>
<tr>
<td></td>
<td>Technician</td>
<td>72.5%</td>
<td>6.1%</td>
</tr>
</tbody>
</table>

Table 4. Duration of Interruptions by Phase and Time of Day and Job Position

<table>
<thead>
<tr>
<th>Phase</th>
<th>Time of Day</th>
<th>Duration of Interruption shown as Mean (SD) in minutes</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-CPOE</td>
<td>Morning</td>
<td>0.8 (0.64)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Afternoon</td>
<td>1.1 (1.31)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evening</td>
<td>1.5 (1.27)</td>
<td></td>
</tr>
<tr>
<td>Immediate Post-CPOE</td>
<td>Morning</td>
<td>1.2 (1.22)</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Afternoon</td>
<td>1.9 (2.05)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evening</td>
<td>2.1 (2.15)</td>
<td></td>
</tr>
<tr>
<td>Later Post-CPOE</td>
<td>Morning</td>
<td>1.0 (0.95)</td>
<td>Not significant</td>
</tr>
<tr>
<td></td>
<td>Afternoon</td>
<td>1.2 (1.03)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evening</td>
<td>1.6 (1.76)</td>
<td></td>
</tr>
</tbody>
</table>

711
Duration was analyzed using regression, where the durations were converted to a log scale. Based on data collection, window, door and hand-on task interruptions were aggregated to be listed as “Other.” This is both because, of the primary interest in incoming phone calls, outgoing phone calls and discussion interruptions and of the small frequency of these “other” interruptions not all of these occur in all stratifications, for instance, there are no morning hands-on tasks or evening door interruptions.

Given the sample size, the duration was analyzed using t-test. Overall, duration of interruptions showed several significant relationships. Durations of interruptions involving pharmacist are significantly longer than interruptions involving technicians (p<0.0001). Table 5 shows “Technician” as the reference for “Job Position” and the estimate is positive. Similarly, the durations of interruptions were significantly different during various phases of CPOE (p=0.0006). Specifically, the duration of interruptions in the pre-CPOE phase were significantly shorter than the interruptions immediately post-CPOE (P=0.0001) and later post-CPOE (P=0.0006). Furthermore, durations of interruptions immediately post-CPOE were found to be significantly longer than interruptions later post-CPOE (p=0.0012). Stress type by itself was not significant with regard to the duration of interruptions. However, stress type along with the time of day have an interaction effect and significantly affect the duration of the interruption (p=0.0056). Please refer to Figure 1 for the trend in interruption duration of stress type by time of day. Overall, durations shows an increasing trend as the day progresses with lot of variation in the afternoon. Further, there are different patterns for different type of interruptions. Incoming calls fairly stay the same (with a slight increase) from morning to afternoon and drastically increase into the evening. On the other hand, outgoing calls increase rapidly from morning to afternoon and stay approximately the same into the evening. And, discussions have a gradual increase through the day. Other interruptions show drastic changes, with the least duration in the morning to the maximum in the afternoon and the least in the evening.

Figure 1. Trend in interruption duration of stress type by time of day (Incoming and Outgoing stressors refer to respective phone calls as stressors)
<table>
<thead>
<tr>
<th>Variables</th>
<th>Est.</th>
<th>SE</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.3092</td>
<td>0.0687</td>
<td>-4.50</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pharmacist (REF: Technician)</td>
<td>0.3963</td>
<td>0.0756</td>
<td>5.25</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Stressor Type (REF: Others)</td>
<td>0.6763</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incoming phone calls</td>
<td>0.0393</td>
<td>0.0596</td>
<td>0.07</td>
<td>0.5097</td>
</tr>
<tr>
<td>Outgoing phone calls</td>
<td>0.0007</td>
<td>0.0992</td>
<td>0.01</td>
<td>0.9943</td>
</tr>
<tr>
<td>Discussion</td>
<td>-0.0988</td>
<td>0.0905</td>
<td>-1.09</td>
<td>0.2758</td>
</tr>
<tr>
<td>Time (REF: Night (After 5PM))</td>
<td></td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morning (Before 11AM)</td>
<td>-0.3357</td>
<td>0.0707</td>
<td>-4.75</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Afternoon (11AM - 5PM)</td>
<td>0.1479</td>
<td>0.0701</td>
<td>2.11</td>
<td>0.0354</td>
</tr>
<tr>
<td>CPOE (Later Post-CPOE)</td>
<td></td>
<td>0.0006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-CPOE</td>
<td>-0.1670</td>
<td>0.0487</td>
<td>-3.45</td>
<td>0.0006</td>
</tr>
<tr>
<td>Immediate Post-CPOE</td>
<td>0.1760</td>
<td>0.0542</td>
<td>3.25</td>
<td>0.0012</td>
</tr>
<tr>
<td>Stress Type x Time (REF: Other Stress &amp; Night (After 5PM))</td>
<td>0.0056</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morning (Before 11AM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incoming phone calls</td>
<td>0.1042</td>
<td>0.0831</td>
<td>1.25</td>
<td>0.2102</td>
</tr>
<tr>
<td>Outgoing phone calls</td>
<td>0.0974</td>
<td>0.1473</td>
<td>0.66</td>
<td>0.5088</td>
</tr>
<tr>
<td>Discussion</td>
<td>0.2032</td>
<td>0.1220</td>
<td>1.67</td>
<td>0.0965</td>
</tr>
<tr>
<td>Afternoon (11AM - 5PM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incoming phone calls</td>
<td>-0.2329</td>
<td>0.0812</td>
<td>-2.87</td>
<td>0.0004</td>
</tr>
<tr>
<td>Outgoing phone calls</td>
<td>-0.0677</td>
<td>0.1373</td>
<td>-0.49</td>
<td>0.6222</td>
</tr>
<tr>
<td>Discussion</td>
<td>-0.1292</td>
<td>0.1221</td>
<td>-1.06</td>
<td>0.2908</td>
</tr>
</tbody>
</table>

CPOE interruptions:
Following these results, the interruption description column was also analyzed. There were a total of 22 interruptions in the immediate post and two interruptions in the later post phases respectively, that were related to CPOE. These were analyzed for differences in types and duration of interruptions. There was a significant difference (p-value=0.0002) in immediate post-CPOE phase between CPOE interruptions and non-CPOE interruptions. Discussions tend to be almost six times more likely with CPOE interruptions. Duration was significantly different (p-value=0.003) for immediate post-CPOE phase. CPOE interruptions are almost twice the length of non-CPOE interruptions.

Discussion
Hospitals have been identified as interruptive environments\(^{28}\), and as can be seen from our study, pharmacists and technicians spend almost 25 percent of their time on interruptions. This study focused primarily on the effect of technology implementation on interruptions in inpatient pharmacy. Overall, CPOE implementation disrupted order processing operations, and specifically had an impact on the type of interruptions and duration, but insufficient evidence exists with regard to the impact on the frequency of interruptions. CPOE implementation, by itself, had an impact only on the duration of the interruptions. With the type of day, it also had an impact on the types of interruptions. And, along with time of day and job position, they had an impact on both the types and the duration of interruption. Specifically, the types of interruptions were different, a) by type of day for pre- and later post-CPOE phases; b) by time of day for pre-CPOE phase; and c) by job position for all three phases. The duration of interruptions was different a) by time of day during pre and immediate post-CPOE phases; and b) by job position for immediate post and later post-CPOE phases.

Previous research has found that workers will adapt and use work-arounds to cope with disruptions in work flow following implementation of CPOE\(^{29, 30}\). Phone calls between providers have been used as work-arounds, which may have implications for changes in interruptions following implementation of health information technologies like those trends found in the current study\(^{29}\). We confirm this by noting that almost 70 percent of the interruptions were incoming phone calls and almost 80 percent are all phone-related interruptions.
CPOE interruptions in the immediate post-CPOE phase were almost twice the length of other interruptions, and tend to be six times more likely to lead to discussions. This is consistent with previous work that has shown that computer users prefer to seek guidance or help from peers or local experts instead of printed manuals or other computer-based documentation sources. Thus, post implementation of CPOE, we may see increases in interruption duration as clinicians are consulting the pharmacy not only for their normal questions, but also for technical support and guidance. This has major implications for work flow, since discussion type interruptions have a multiplier effect since they involve interruptions for other pharmacy employees. This increase in duration in the immediate post-CPOE phase may be attributed to “learning curve” phenomenon.

CPOE has been shown to have a significant impact on work processes across medical professions (including nurses, physicians, pharmacists and technicians) by reassigning tasks, reallocating decision-making or expertise, and redistribution of power. We found differences in interruptions based on job position (between pharmacists and technicians). Pharmacists more likely to have discussions and outgoing phone calls, and technician are more likely to have window interruptions. Further, after CPOE implementation, the duration increases considerably more for pharmacists, and the likelihood of discussion interruptions for pharmacists also increases. These highlight that technology implementation may be more disruptive to pharmacists than technicians. Hence, in a technology implemented era, job responsibilities and work flow need to be revisited. Further, as pharmacists assume clinical/rounding pharmacist roles, work redistribution response to interruptions needs to be further studied.

This study considered outgoing phone calls as an interruption disrupting smooth flow of operations. There is a potential argument to consider outgoing phone calls as reducing risk of error and thereby enhancing patient safety. Such a study in the future can focus on distinguishing risk-reducing and risk-increasing interruptions. Our research is based on one hospital and hence restricts its generalizability. The focus is on inpatient pharmacy and does not apply to commercial pharmacy. However, where possible, we link our work to previous research on interruptions in general and shed more light on pharmacy interruptions. The study included sampling from specific times of day; there were no night sessions. Thus, there may be different interruption patterns during time periods not included in the study. However, sessions were scheduled to cover a range of work demand periods across all days of the week and the highest work demand periods of the day. Such targeted observational studies with similar sampling methods have been successfully used to quantify system and worker responses, as well as changes to processes, during naturally occurring changes in an organization, such as the implementation of health information technologies.

Conclusion

Pharmacy employees spend almost a quarter of their time on interruptions, and pharmacists have considerably longer interruptions than technicians during all phases related to CPOE implementation. Immediately after CPOE implementation, durations tend to be one-and-a-half times longer than before, and CPOE interruptions are almost twice as long as non-CPOE interruptions. CPOE implementation does not affect the frequency of interruptions. Our research stresses on the effects of technology implementation and the significant differences in interruptions due to other related factors. Next steps should involve study of origin of interruptions and focus on redesign of work processes and responsibilities of pharmacists and technicians. Further, as the pharmacists’ roles evolve to assume clinical responsibilities as a ‘rounding pharmacist’, changes in the effects on interruptions should be revisited.

References


10. Sobek D, Jimmerson C, editors. Applying the Toyota Production System to a Hospital Pharmacy. Industrial Engineering Research Conference; 2003; Portland, OR.


**aer2vec: Distributed Representations of Adverse Event Reporting System Data as a Means to Identify Drug/Side-Effect Associations**

Jake Portanova, BA, BS¹, Nathan Murray, BA², Justin Mower, PhD², Devika Subramanian, PhD², Trevor Cohen, MBChB, PhD¹

¹University of Washington, Seattle, WA, ²Rice University, Houston, TX, ³Hofstra University, East Garden City, New York

**Abstract**

Adverse event report (AER) data are a key source of signal for post marketing drug surveillance. The standard methodology to analyze AER data applies disproportionality metrics, which estimate the strength of drug/side-effect associations from discrete counts of their occurrence at report level. However, in other domains, improvements in predictive modeling accuracy have been obtained through representation learning, where discrete features are replaced by distributed representations learned from unlabeled data. This paper describes aer2vec, a novel representational approach for AER data in which concept embeddings emerge from neural networks trained to predict drug/side-effect co-occurrence. Trained models are evaluated for their utility in identifying drug/side-effect relationships, with improvements over disproportionality metrics in most cases. In addition, we evaluate the utility of an otherwise-untapped resource in the Food and Drug Administration (FDA) AER system – reporter designations of suspected causality – and find that incorporating this information enhances performance of all models evaluated.

**Introduction**

The need for post-marketing surveillance of the unintended effects of pharmaceutical products has been repeatedly highlighted by the withdrawal of drugs on account of previously undetected serious side effects¹. Notoriously, the widely utilized Cox-2 inhibitor Vioxx (rofecoxib) was removed from market five years after release as it was shown to significantly increase the risk of myocardial infarction, and was estimated to have caused between 88,000 and 140,000 coronary cardiac events in the United States alone while on market²,³. More recently proton pump inhibitors, many of which are available without prescription, have been shown to significantly increase risk of bone fracture as well as chronic kidney disease⁴,⁵. These findings are not atypical – between the years of 2001 and 2010, nearly one third of drugs approved by the US Food and Drug Administration (FDA) had a subsequent safety event in the form of a label change, safety communication or withdrawal⁶. The median time from drug release to detection of such events was 4.2 years⁶. The morbidity and mortality caused by previously undetected drug side effects could be mitigated by earlier detection, as could the societal costs of such adverse events – estimated at 3.5 billion dollars in 2006⁷. One solution to the inherent shortcomings of clinical trials in detecting adverse drug events (ADEs) is improving their identification after release to market. Consequently, post-marketing surveillance through pharmacovigilance, defined as “the study of the safety of marketed drugs under the practical conditions of clinical use in large communities”, is an essential component of drug safety⁸.

Traditionally, such safety surveillance has involved the analysis of reports of suspected ADEs submitted by healthcare practitioners, pharmaceutical companies and patients. In the United States the FDA maintains the FDA Adverse Event Reporting System (FAERS), providing a database of ADE reports from as early as 1969⁹. Adverse events in the AERS are reported by healthcare professionals, consumers, and pharmaceutical companies. Each report includes one or more adverse events that appear to be associated with the administration of a drug as well as other drugs prescribed to the patient concerned and their therapeutic indications. Of importance to the current research, reporters have the opportunity to indicate which of a set of prescribed drugs they suspect caused the ADE under consideration, by designating these drugs as primary (presumed cause) or secondary (potential cause) suspects. Large numbers of reports exist in this repository, with over a million received in the year 2014 alone¹. Consequently, automated methods of analysis are a prerequisite to the identification of actionable safety signals.
An important component of post-marketing drug surveillance is the identification of statistically significant drug/side-effect association, termed “signal detection”, and considerable research has been devoted to the development and evaluation of methods for this purpose. In order to identify meaningful associations from large and unlabeled data sets such as AERS reports, data mining techniques known as Signal Detection Algorithms (SDAs) are employed. SDAs can be subdivided into two main classes, disproportionality analysis (DPA) and multivariate modeling. DPA methods quantify the extent to which drugs and side-effects are reported together beyond what would be expected by chance. For a review of DPA methodologies, we refer the interested reader to Bate et al. Multivariate modeling can mitigate for confounding polypharmacy variables as well as the lack of a quantitative drug prescription frequency. SDAs are the most frequently utilized methods used in analyzing AERS data, and they are well documented in the post-marketing drug surveillance literature.

While SDAs have shown their utility as a means to identify safety signals from adverse event reports, they are not without limitations. In particular, they lack the capacity to draw connections between similar drugs (e.g. the entire family of selective cox-2 inhibitors) and related side-effects (e.g. myocardial infarction and other cardiac events such as stroke) as a means to enhance the strength of safety signals for relatively rare events. This situation is analogous to recent developments in natural language processing (NLP), where discrete ‘one-hot’ vector representations of words have been largely superseded by distributed vector representations, which are learned from a large unlabeled corpus such that words that occur in similar contexts have similar vectors. In the current paper we adapt skip-gram neural embeddings, a widely used representation learning technique in NLP embodied in the popular word2vec software package, to the task of representing drugs and side-effects appearing in FAERS data and evaluate their utility as a means to detect safety signals.

Specifically, this paper describes aer2vec, a novel representational approach for adverse event report data in which a neural network is trained to predict drug/side-effect co-occurrence events. The trained model is evaluated for its utility as a means to identify causal drug/side-effect relationships. Our approach leverages methods of distributional semantics to represent the AERS database as a vector space. Distributional semantics methods - such as the neural embeddings implemented by word2vec - attempt to model the semantic similarity and relatedness between words. These methods are based on the distributional hypothesis which states that words that occur in similar contexts tend to have similar meanings. A broad range of methodological approaches have been applied to learn word representations from text (for reviews see). A recent trend involves the application of neural-probabilistic models, such as the skipgram and continuous-bag-of-words architectures embodied in the popular word2vec and fastText software packages. As implied by the term “neural-probabilistic”, these models are trained to predict the occurrence of a context word given an observed term. Although generally utilized during training only, this probabilistic aspect of the model can be used to recover learned probabilities for observing one word in the context of another. It is this aspect of neural-probabilistic models that we adapt to represent AERS data in the current work.

Our primary hypothesis in conducting this work was that, at least in some cases, the capacity of distributional semantics models to generalize between similar drugs and side-effects may lead to improved performance in the task of identifying drug/side-effect relationships. A secondary hypothesis was that restricting the data considered by our models, and perhaps baseline disproportionality metrics also, to designations of primary and/or secondary suspect may improve their performance.

Methods

Disproportionality metrics: We compare the performance of our models to two widely-used disproportionality metrics, the Proportional Reporting Ratio (PRR) and the Reporting Odds Ratio (ROR). Disproportionality metrics are derived from 2x2 contingency tables (Table 1) constructed using report level statistics:

<table>
<thead>
<tr>
<th></th>
<th>ADE of Interest</th>
<th>Other ADE’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug of Interest</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Other Drugs</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

718
From this table, we derive two disproportionality metrics, the PRR and the ROR. The PRR estimates the probability of an ADE given a drug divided by the probability of this ADE without the drug, or \( P(\text{ADE} | \text{drug}) / P(\text{ADE} | \neg \text{drug}) \). This probability can be calculated from the 2x2 table as follows:

\[
PRR = \frac{a}{a+b} / \frac{c}{c+d}
\]

In statistical terms, the ROR is the ratio of odds of an ADE occurring vs not occurring given a drug to the odds of this ADE occurring vs not occurring given other drugs. It is calculated from the 2x2 table as follows:

\[
ROR = \frac{a}{b} / \frac{c}{d}
\]

These disproportionality metrics provide a baseline for our evaluation.

Skipgram-with-negative-sampling: The skipgram-with-negative-sampling (SGNS) algorithm of Mikolov and his colleagues trains a shallow neural network to predict the probability of a context word \( c \) occurring nearby to an observed word \( w \), \( P(c | w) \). For example, given the sliding window “adverse [drug] events”, the model would be trained to optimize \( P(\text{adverse} | \text{drug}) \) and \( P(\text{events} | \text{drug}) \). While it would in theory be possible to train this model using a softmax objective in which all other unique terms in the corpus are considered as counterexamples to the observed context term in every window, this would be computationally intractable. The SGNS algorithm provides a practical way to train neural embeddings by instead considering as counterexamples a small number (usually 5-15) of negative samples, randomly drawn words that probably do not occur with the focus word within a sliding window. More formally, the optimization objective of the SGNS algorithm can be expressed as follows:

\[
\arg\max_{\theta} \sum_{(w,c) \in D} \log \sigma(v_c, v_w) + \sum_{(w,c') \in D'} \log \sigma(-v_{c'}, v_w)
\]

where \( w \) are observed words in a set of documents \( D \), and \( c \) and \( c' \) are context terms that occur with these observed terms in a sliding window or are randomly drawn as counterexamples respectively. \( v_w \) and \( v_c \) are vector representations of these terms, specifically the input weights \( (v_w) \) and output weights \( (V_c) \) of the neural network for each term in the vocabulary. \( \sigma \) is the sigmoid function, which converts the scalar product between these vectors into a value between one and zero which can be interpreted probabilistically. Training in SGNS occurs through stochastic gradient descent with a linearly decreasing learning rate. The algorithm has several hyper-parameters that have been shown to influence performance across tasks. Of importance for the current paper, the number of negative sample terms drawn as counterexamples to each observed term is a parameter of the model, and these terms are drawn with a probability derived from the frequency, \( f \), with which they appear in the corpus (i.e. \( f = \text{count} / \text{total number of non-unique terms} \)). Another hyper-parameter setting concerns subsampling – ignoring terms occurring above a predetermined frequency threshold \( t \) with probability \( 1 - \frac{2}{\sqrt{f}} \). With SGNS, the input weights are usually retained as word embeddings and used in downstream tasks, and the output weights are discarded. However, \( P(c | w) \) can be estimated by retaining these weights and calculating \( \sigma(v_c, v_w) \).

aer2vec: With aer2vec, we adapt the SGNS algorithm to estimate probabilities for drug/side-effect relations from AERS data. We derive two neural architectures, aer2vec + which estimates \( P(\text{drug} | \text{ADE}) \) and aer2vec - which estimates \( P(\text{ADE} | \text{drug}) \). Our motivation in developing distinct architectures for each of these estimates was to ensure that only information pertinent to an estimate of interest is encoded by the architecture responsible for it, eliminating one channel through which noise might be introduced. In addition we wished to determine which of these two possible neural-probabilistic estimates is of greater utility for identification of drug/ADE relationships.
These architectures are illustrated in Figure 1, which shows a simplified aer2vec architectures for 8 side effects and 10 drugs. As is the case with the original skipgram architecture, each observed term (in this case a drug) is connected to a hidden layer by input weights, which are in turn connected to a predicted term by output weights. While the output weights are usually discarded when generating word embeddings, we retain them to facilitate predicting \( \text{P} \text{drug} | \text{ADE} \) and \( \text{P} \text{ADE} | \text{drug} \) in aer2vec+ and aer2vec- respectively. When trained with negative sampling, the aer2vec architectures have the following optimization objectives:

\[
\text{aer2vec}^+: \arg \max_\theta \sum_{(\text{ADE}, \text{drug}) \in R} \log \sigma(v_{\text{drug}}^i, v_{\text{ADE}}^j) + \sum_{(\text{ADE}, \text{drug'}) \in R'} \log \sigma(-v_{\text{drug}}^i, v_{\text{ADE}}^j)
\]

\[
\text{aer2vec}^-: \arg \max_\theta \sum_{(\text{drug}, \text{ADE}) \in R} \log \sigma(v_{\text{ADE}}^i, v_{\text{drug}}^j) + \sum_{(\text{drug'}, \text{ADE'}) \in R'} \log \sigma(-v_{\text{ADE}}^i, v_{\text{drug}}^j)
\]

This is similar to SGNS with word windows, with the exceptions that (1) co-occurrence occurs at the level of a report (\( R \)) rather than a sliding window; and (2) drugs and ADEs have either input weight vectors or output weight vectors depending on the model, whereas words in neural word embeddings have both input and output weights. With aer2vec, both sets of weights are retained, permitting estimation of both \( \text{P} \text{drug} | \text{ADE} \) with aer2vec+, and \( \text{P} \text{ADE} | \text{drug} \) with aer2vec-, as \( \sigma(v_{\text{drug}}^i, v_{\text{ADE}}^j) \).

**Evaluation**

**AER data:** As data for all models, we used a standardized version of the FAERS data set released by Banda et al\(^2\). This data set contains standardized versions of the FAERS data for the years 2004 to 2013. Standardization here indicates that drugs are mapped to RxNorm concepts, and the side-effects to SNOMED-CT\(^2\). We extracted three component data sets. First, we generated an unconstrained data set (FULL), including all the drugs that were mapped to RxNorm and side effects that occur in the “pt” field of any report with them. We then use the role code variable to produce two subsets of this data set. The role codes we used were primary suspect (PS) and secondary suspect (SS). These terms denote whether a drug is the primary or secondary suspect for a side-effect, and we leveraged these to produce data subsets consisting of only drugs and side-effects in primary suspect relationships (PS), and the disjunction of primary and secondary suspect relationships (PS-SS).

**aer2vec models:** For each of the three dataset configurations, we generated aer2vec+ and aer2vec- models. All models used 500-dimensional vectors and were trained for 25 iterations across the data with five negative samples per positive example, and without subsampling of frequently occurring concepts. We did not attempt to optimize these hyperparameters in our initial experiments. During training we encoded all of the drug and ADE terms meeting the relevant constraints (PS/PS-SS/FULL), aside from a stoplist of 28 ADE terms (such as “NA”, “test”, “error”) that were identified as lacking semantic content upon inspection of the data.
**Evaluation sets:** A number of research groups have released reference standards that provide a uniform testing ground for pharmacovigilance systems. In some cases, these are manually curated reference sets of positive and negative relationships between drugs and potential ADEs\textsuperscript{23,24}. For example, the widely-used OMOP reference standard produced by Ryan and his colleagues consists of 165 positive and 234 negative examples of drug/ADE relationships, spanning four serious side-effects (such as renal failure), and developed through extensive manual review of the literature and other sources\textsuperscript{23}. These sets have been extensively curated, so there is reason to believe they are largely accurate (for positive examples in particular – some amendments have been suggested for negative controls). We evaluated aer2vec for its ability to distinguish between positive and negative controls in the reference standards produced by Ryan \textit{et al} (henceforth, OMOP) and Coloma \textit{et al} (henceforth EU-ADR) containing 399 and 94 examples respectively\textsuperscript{23,24}. For both disproportionality metrics and aer2vec models, a small number of examples were eliminated from the FULL (n=1) and PS/PSSS (n=6) configurations of the OMOP set respectively. These examples concerned the drugs “olmesartan medoxomil” (all configurations), “endopeptidases” and “alatrofloxacin” (PS/PSSS configurations), which are not present in the source data. All examples from the EU-ADR set (n=94) were retained.

**Quantitative evaluation:** For each drug/event pair in the two reference standards (OMOP/EU-ADR) under each of the three configurations (PS/SS/FULL), we calculated disproportionality metrics (PRR/ROR), as well as $P(\text{drug}/\text{ADE})$ and $P(\text{ADE}/\text{drug})$ with aer2vec+ and aer2vec- respectively. These scores were used to estimate the area under receiver operator characteristic curve (AUROC) and area under the precision recall curve (AUPRC) to provide a basis for comparison. In order to account for variance in the stochastic initialization of the aer2vec embedding weights, we retrained each aer2vec model 10 times, and report mean area under curve (AUC) across these runs.

**Qualitative evaluation:** In order to assess the likely utility of the vector embeddings generated during the course of training our models for downstream supervised machine learning, we include some examples of nearest neighbor searches amongst the input and output embeddings for the best-performing model.

**Hyperparameter settings:** Subsequent to our initial experiments, we explored the influence of two aer2vec+ hyperparameters on model performance. We repeated our initial experiments with the PS and FULL configurations at dimensionality of 100, 250, 500 (the original setting) and 1,000 dimensions; and at 100 dimensions with subsampling thresholds of $10^3$, $10^4$, $10^5$ and without subsampling (the original setting). For these experiments we used 5 (rather than the original 25) epochs of training.

**Results:**

**Quantitative evaluation:** Disproportionality metric results for all three versions of the data sets across both reference standards, are shown in Table 2. These results provide points of comparison for aer2vec. As anticipated (see for example Waller \textit{et al} \textsuperscript{23}), these metrics exhibit similar performance. Of note, both baseline metrics show an increase in performance when reducing the data set by role code. In the OMOP set, restricting to the PS-SS leads to best performance with AUROC’s of .744 for both PRR and ROR. In the EU-ADR set, best performance is obtained with the PS constraint with AUROC’s of .935 for both methods. Across both reference sets and disproportionality metrics, there is an absolute AUROC increase of approximately 10% when we reduce the full data set to the primary suspects.

<table>
<thead>
<tr>
<th></th>
<th>OMOP</th>
<th>EU-ADR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FULL</td>
<td>PS-SS</td>
</tr>
<tr>
<td><strong>PRR</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUROC</td>
<td>0.646</td>
<td>0.744</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.580</td>
<td>0.694</td>
</tr>
<tr>
<td><strong>ROR</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUROC</td>
<td>0.646</td>
<td>0.744</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.579</td>
<td>0.691</td>
</tr>
</tbody>
</table>

**aer2vec:** Results for the two aer2vec models are shown in Table 3. These results show improvement over disproportionality metrics for both reference sets in every configuration (FULL,PS-SS and PS) by both metrics of evaluation. As was the case with the disproportionality metrics, the results show an improvement in AUC when reducing the data set by role code, with an absolute AUC increase of approximately 10% when we constrain the
FAERS data to primary suspects. While aer2vec+ and aer2vec- perform similarly to one another on the EU-ADR set, the aer2vec+ model performs considerably better on the larger OMOP set. As was the case with disproportionality metrics, performance improves by around 10% (absolute AUROC) when restricting data to primary suspect relationships only. In addition, the best aer2vec results exceed the best disproportionality metric result on both data sets, with substantive improvements in performance across all configurations with the OMOP set in particular. As indicated by the confidence intervals of these mean values, performance was remarkably consistent across iterations. Across these ten iterations, all differences in performance between PS, PS-SS and FULL configurations of the same model were statistically significant by an unpaired t-test, as were all differences between aer2vec+ and aer2vec- models in the same configuration, aside from the sole case of the PS-SS configuration with the EU-ADR set.

Table 3: aer2vec results. Mean AUROC/AUPRC ±1.96*STD across 10 repeated runs. PS-SS= primary and secondary suspects. PS=primary suspects. Best results for each metric on each reference set are in boldface. The best baseline model (ROR) is shown in the middle two rows for the purpose of comparison.

<table>
<thead>
<tr>
<th></th>
<th>OMOP</th>
<th>EU-ADR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FULL PS-SS PS</td>
<td>FULL PS-SS PS</td>
</tr>
<tr>
<td>AER2VEC+</td>
<td>AUROC: 0.727 ± 0.801</td>
<td>0.839 ± 0.849 ± 0.937 ± 0.952</td>
</tr>
<tr>
<td></td>
<td>±0.016 ±0.008 ±0.007 ±0.024 ±0.006 ±0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUPRC: 0.679 ± 0.759</td>
<td>0.809 ± 0.848 ± 0.946 ± 0.960</td>
</tr>
<tr>
<td></td>
<td>±0.019 ±0.008 ±0.007 ±0.033 ±0.006 ±0.003</td>
<td></td>
</tr>
<tr>
<td>ROR</td>
<td>AUROC: 0.646 ± 0.744</td>
<td>0.742 ± 0.843 ± 0.919 ± 0.935</td>
</tr>
<tr>
<td></td>
<td>±0.015 ±0.010 ±0.007 ±0.015 ±0.007 ±0.007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUPRC: 0.579 ± 0.691</td>
<td>0.723 ± 0.840 ± 0.932 ± 0.941</td>
</tr>
<tr>
<td></td>
<td>±0.017 ±0.011 ±0.008 ±0.019 ±0.008 ±0.007</td>
<td></td>
</tr>
</tbody>
</table>

Table 4 illustrates three ways in which the trained embedding space (in this case aer2vec+ PS) can be interrogated. The first two results columns involve comparing the same types of entities – side effects (in this case input weight embeddings) or drugs (in this case output weight embeddings). Furthermore, it is possible to recover predictions for observing a particular drug given a side effect has been reported as shown in the third results column. As indicated in the accompanying comments, the results are interpretable in many cases and suggest the resulting embeddings may be of utility for downstream supervised learning models on account of the similarities they capture.

Table 4: Nearest neighbors in aer2vec+ PS space. IW=input weights, OW=output weights

| Cosine to IW(dystonia) | Cosine to OW(clozapine) | P(drug|acute myocardial infarction) |
|------------------------|-------------------------|--------------------------------|
| 1.00 dystonia          | 1.00 clozapine*          | 0.54 rofecoxib*                 |
| 0.67 tardive dyskinesia| 0.41 trimeprazine*       | 0.28 rosiglitazone*             |
| 0.65 extrapyramidal disorder | 0.39 clorazepanone | 0.23 ticagrelor                  |
| 0.59 akathisia         | 0.38 piracetam           | 0.15 prasugrel                   |
| 0.56 meige's syndrome  | 0.37 moclobemide         | 0.11 bivalirudin                |
| 0.55 dyskinesia        | 0.37 nitrazepam          | 0.08 alvimopan                  |
| 0.55 oculogyration     | 0.36 loxapine*           | 0.07 abacavir/lamivudine         |
| 0.55 respiratory dyskinesia | 0.36 perazine* | 0.04 tirosiban                    |
| 0.53 oculogyric crisis | 0.35 amisulpride*        | 0.03 valdecoxib*                |
| 0.52 neuroleptic malignant syndrome | 0.34 sulphiride* | 0.01 ziconotide_acetate          |

Other mostly movement related side effects caused by antipsychotic drugs, (*): drugs sharing rare ADEs: except Meige's syndrome, a dystonia toxic epidermal necrolysis of unknown etiology (surprising to (clorazepanone); agranulocytosis (loxapine). Removed from market for cardiac safety concerns (*). Myocardial infarction is a known side effect of abacavir. Some other results (anticoagulants and platelet inhibitors) suggest reporter misattribution of suspicion.
These results also give some insight into the possible mechanisms underlying aer2vec’s improvements in performance over established disproportionality metrics, as the estimation of \( P(\text{drug|ADE}) \) (or vice-versa) will be influenced by similarities between drug and side-effect terms. For example, \( P(\text{drug|tardive dyskinesia}) \) will be elevated for a drug reported with the term “extrapyramidal disorder”, which is a useful generalization in this case because of the hyponymic relationship between these side effects (tardive dyskinesia is a type of extrapyramidal disorder). More broadly, one would anticipate the model inferring that drugs that share some side effects may share others.

**Hyperparameter settings:** Evaluation of performance across different hyperparameter settings reveals that subsampling impairs performance (Figure 2, left) and consistency across a range of dimensions (Figure 2, right). Best performance across these settings is shown in Table 5, with improvements in OMOP set performance in particular.

**Figure 2:** Mean AUROC \((n=10)\) with 95% CI at different subsampling thresholds (left; ‘-’ indicates no subsampling) and dimensionalities (right). 0.5 baseline indicates performance with random ordering.

**Table 5:** Best aer2vec+ performance across hyperparameter settings (100 dimensions, no subsampling)

<table>
<thead>
<tr>
<th>Source Set</th>
<th>OMOP FULL</th>
<th>OMOP PS</th>
<th>EU-ADR FULL</th>
<th>EU-ADR PS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AER2VEC+</strong></td>
<td>AUROC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.761±0.013</td>
<td>0.880±0.004</td>
<td>0.861±0.026</td>
<td>0.960±0.006</td>
</tr>
<tr>
<td></td>
<td>AUPRC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.712±0.021</td>
<td>0.844±0.007</td>
<td>0.863±0.038</td>
<td>0.968±0.008</td>
</tr>
</tbody>
</table>

**Discussion**

Our results demonstrate improved signal detection from the FAERS data when applying an adaptation of the SGNS algorithm to represent observational data, providing a novel methodology for the identification of drug/side-effect associations. Across both reference standards and three different configurations of our source dataset, \( P(\text{drug|ADE}) \) estimates produced by the aer2vec+ model consistently result in a higher AUC than was obtained by either of the two disproportionality metrics evaluated. The aer2vec- model also outperformed disproportionality metrics in most but not all cases. Furthermore, we found that the role code information in the FAERS data can be used to subset these data in a way that improves performance. Lastly, we discovered that aer2vec+ outperforms aer2vec- in all cases. These findings have implications for the analysis of adverse event report data.

Our primary finding is that aer2vec+ outperforms the standard disproportionality metrics across two reference sets and all dataset configurations, with improvements in performance of up to 10% in absolute AUROC on the larger and more challenging OMOP set, and smaller but still substantial improvements in performance on the EU-ADR set. This supports our initial hypothesis that distributed representations (concept embeddings) derived from AERS data may be of value for pharmacovigilance. Qualitative analysis of neighboring embeddings suggests that trained embeddings capture regularities in the data. Concepts occurring in similar AER contexts have similar embeddings, permitting our
models to generalize across related drugs and ADEs, suggesting a mechanism underlying their enhanced performance. In the current work, we applied these embeddings by recapitulating the probabilistic estimates used during the course of training. However, the benefits of pre-trained distributed representations are often best demonstrated in downstream supervised learning. In our previous work we have shown that literature-derived embeddings used as a basis for supervised machine learning produce state-of-the-art performance (AUROC=.96) on the OMOP reference set\textsuperscript{26,27}. A logical next step will involve evaluating the utility of aer2vec embeddings as a complement to literature-derived embeddings for supervised machine learning. Intriguingly, best performance with aer2vec on the EU-ADR set exceeds that obtained with supervised models - both in our work\textsuperscript{26,27} and when classifiers were trained on manually engineered features derived from a range of sources including the biomedical literature, FAERS data and pharmaceutical product labels\textsuperscript{38}. One reason for this may be the limited number of training examples per side-effect available relative to the OMOP set, where recapitulating aer2vec’s neural-probabilistic learning objective does not in and of itself lead to state-of-the-art performance. In addition, prior work leveraging side-effect patterns in AERS data for drug repurposing suggests aer2vec embeddings that represent such patterns may be leveraged for this purpose also\textsuperscript{29}.

A second notable finding is that constraining the data set in accordance with human judgment of suspected causality enhances performance. The role code variable in the FAERS data set provides valuable information that has generally not been used in prior published work, and leveraging this information to constrain the occurrence events in the data resulted in substantive improvements in performance for all models evaluated. It seems likely that the mechanism in this case involves reducing the influence of confounding variables, providing a useful addition to the armament of methods that have been deployed for this purpose in pharmacovigilance\textsuperscript{30,31,32}. However, incorporation of human judgment into statistical models introduces a degree of subjectivity that warrants further discussion. One might argue that attempts to leverage these designations of suspicion would accentuate documented biases in spontaneous reporting, such as notoriety bias - a tendency toward increased reporting of adverse events that have recently appeared in safety alerts\textsuperscript{33}. A more pragmatic perspective might be to consider the cumulative independent judgments of a diverse group of adverse event reporters – who are in some cases domain experts and have access to information concerning temporality and other factors that are not explicit in reporting data alone – as exemplifying the “wisdom of crowds”\textsuperscript{34}. At a minimum, the utility of reporter designations of suspected causality as a means to improve predictive modeling performance across multiple models and two reference standards suggests a need for further investigation into the circumstances in which this information is of value for post-marketing drug surveillance.

In addition, we find that aer2vec+ performs better than aer2vec-. While both approaches outperform the disproportionality metrics in most cases, aer2vec+ is consistently better than aer2vec-, with better mean AUC scores for aer2vec+ across both reference sets in all configurations. We reason that predicting the drug given the side-effect is the more suitable approach to identifying drug/side-effect associations from AERS data, and that future research should prioritize aer2vec+ as a means to represent these data for pharmacovigilance purposes. Subsequent exploration of model hyperparameters showed subsampling of frequent terms reduced performance, which seems intuitive given the nature of the task. In addition, these experiments revealed improved aer2vec+ performance on the OMOP set in particular at a lower dimensionality than in our original experiments. Optimal performance at lower dimensionality has been observed in other distributional semantic models also, and we refer the interested reader to Yin and Shen’s recent theoretical account of the relationship between dimensionality and task performance in such models\textsuperscript{35}. Code to reproduce our experiments is publicly available\textsuperscript{i}, as are trained models for our primary and best results\textsuperscript{ii}.

Limitations
An important limitation of this study concerns the use of reference sets containing well-established side effects. More stringent evaluations using time-indexed sets are needed to determine performance with emerging side effects\textsuperscript{36}. These models were also created with standardized data. Our methods have yet to be evaluated in the context of raw FAERS data, although one might hypothesize that the capacity for generalization provided by distributed representations would offer greater advantages in the context of data that have not been normalized. Another limitation is that this data set only contains data up to 2015. We plan to address these limitations by training our models using raw FAERS data in future experiments. Finally, we compared aer2vec performance to two disproportionality metrics only. Outperforming these metrics is of practical significance because they are currently in use for regulatory purposes. However in recent work these and other statistical metrics have been evaluated for utility as features for supervised machine learning. In future work we will assess aer2vec’s utility as a basis for downstream machine learning also.

---

\textsuperscript{i} https://github.com/treversee/aer2vec
\textsuperscript{ii} https://zenodo.org/record/3283012
Conclusion

In this paper, we described how aer2vec distributed representations of AER data can be used to acquire signal for drug/side-effect associations, outperforming established disproportionality metrics on two pharmacovigilance reference standards. Considering provider designations of suspected causality resulted in further improvements in performance in both aer2vec and baseline models. As the baseline disproportionality models concerned are in current use, these results have immediate implications for pharmacovigilance practice, with the potential for broad application of the embeddings that result.

Acknowledgments

This work was supported by U.S. National Library of Medicine Grant (R01-LM011563), Robust Inference from Observational Data with Distributed Representations of Conceptual Relations.

References


A High Recall Classifier for Selecting Articles for MEDLINE Indexing

Alastair R. Rae*, PhD, Max E. Savery*, James G. Mork, MSc, Dina Demner-Fushman, MD, PhD
Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD

* Authors contributed equally

Abstract

MEDLINE is the National Library of Medicine’s premier bibliographic database for biomedical literature. A highly valuable feature of the database is that each record is manually indexed with a controlled vocabulary called MeSH. Most MEDLINE journals are indexed cover-to-cover, but there are about 200 selectively indexed journals for which only articles related to biomedicine and life sciences are indexed. In recent years, the selection process has become an increasing burden for indexing staff, and this paper presents a machine learning based system that offers very significant time savings by semi-automating the task. At the core of the system is a high recall classifier for the identification of journal articles that are in-scope for MEDLINE. The system is shown to reduce the number of articles requiring manual review by 54%, equivalent to approximately 40,000 articles per year.

Introduction

MEDLINE® is the U.S. National Library of Medicine’s (NLM) journal citation database. It contains over 25 million citations and is growing at rate of over 900,000 citations per year. The subject scope of MEDLINE is biomedicine and health, broadly defined to encompass the information needs of those working in healthcare and life sciences. A distinctive feature of MEDLINE is that records are manually indexed with NLM Medical Subject Headings (MeSH®). MEDLINE currently covers over 5,200 international journals, and the majority of these journals are indexed cover-to-cover; that is, all articles, substantive editorials, and letters are indexed. About 200 journals are selectively indexed, meaning that only articles related to biomedicine and life sciences are indexed. These selectively indexed journals are typically multidisciplinary journals, such as general science or general chemistry titles.

Between 2000 and 2017, the number of articles from selectively indexed journals has increased rapidly from approximately 13,000 to 78,000, while at the same time the fraction of articles selected for indexing has decreased from about 90% to 25%†. As a consequence, selecting articles for MEDLINE indexing is an increasing burden for NLM indexers, distracting them from their core task of indexing the biomedical literature.

This paper presents a machine learning based system that has been developed to assist indexers with the selection process. The core component of the system is a high recall classifier for the identification of journal articles that are in-scope for MEDLINE. The main inputs to the classifier are the article title and abstract, and the output is a prediction of whether the article is in-scope for MEDLINE. The classifier is used to pre-filter articles that are likely to require MEDLINE indexing and offers significant time-savings by reducing the number of articles requiring manual review. Indexers require the classifier to have close to 100% recall as they do not want to miss in-scope articles.

Related Work

We are not aware of any prior work on selecting biomedical journal articles for indexing, however triage of relevant PubMed articles is often the first step in database curation. An example of such an approach is the TREC 2004 genomics track categorization task2. This task simulated the curation of the Mouse Genome Informatics system and required the triage of articles that contain evidence supporting the assignment of a GO code to a specific gene. The goal was to limit the number of articles sent to human curators for more exhaustive analysis. A similar example is the selection and ranking of articles for the Comparative Toxicogenomics Database (CTD) in the BioCreative evaluation3. The CTD database captures chemical-gene-disease relationships and triage effectiveness was measured by gene, disease, and chemical named entity recognition performance. A different application, that also requires the selection of a subset of PubMed articles, is the identification of scientifically rigorous articles for clinicians practicing

† Statistics computed using the 2018 MEDLINE/PubMed baseline4.
evidence-based medicine. Several machine learning based approaches have been developed to solve this important problem\textsuperscript{[5,6]}. More generally, topic-based text classification has been extensively studied in many domains. A relevant and well-studied topic-based text classification problem is the automatic indexing of biomedical articles using the MeSH vocabulary\textsuperscript{[7,8]}. 

**Methods**

**Dataset**

The dataset is comprised of citation data for MEDLINE articles published in selectively indexed journals before September 2018. It was constructed using a list of selectively indexed journals (and associated start/end years of selective indexing) automatically extracted from the 2018 NLM List of Serials Indexed for Online Users file\textsuperscript{10}. Citation data was downloaded from the 2018 MEDLINE/PubMed annual baseline\textsuperscript{11} and daily update files, and articles from selectively indexed journals were filtered based on their journal and publication date. Articles published in the same year that their journal was selected or deselected for selective indexing were excluded as selective indexing may have started or ended part way through the year.

Articles from selectively indexed journals have typically been reviewed by a single NLM indexer, and whether or not an article has been selected for indexing can be determined from the MEDLINE citation status. Indexed articles are assigned “MEDLINE” status and out-of-scope articles are assigned “PubMed-not-MEDLINE” status. There are 33 journals that are known to have problematic determinations before 2015, and the affected articles were removed from the dataset. The validation and test sets only contain articles published in 2018, as we want to evaluate how the system performs on recent articles. The test set only contains articles from a subset of 132 selectively indexed journals that the indexing team are particularly interested in\textsuperscript{‡}. Some special citation types (Comment, Erratum, Expression of Concern, Republished, Retraction, Update, Reprint, and Patent Summary) are also excluded from the validation and test sets because they follow special indexing rules. The final dataset contains 1.5 million training set articles, 14,346 validation set articles, and 29,833 test set articles. As discussed previously, the fraction of articles selected for indexing has decreased over time: the training set contains articles published before 2018 and has an overall indexing rate of 64%, while the validation and test sets contain articles published in 2018 and have indexing rates of 18%.

**Classifier for In-Scope Biomedical Journal Articles**

The developed classifier combines the predictions of an ensemble of traditional machine learning algorithms and a Convolutional Neural Network (CNN). These two component classifiers are described in detail below.

**Ensemble of Traditional Machine Learning Algorithms**

The ensemble is implemented using scikit-learn\textsuperscript{\textsuperscript{[11]}v0.20.2) and uses averaging to combine the predictions of Bernoulli Naive Bayes, Logistic Regression, Stochastic Gradient Descent, and Random Forest classifiers. The model was trained on 2017 data and the input features are concatenated term frequency inverse document frequency (TF-IDF) representations of the title, abstract, and author affiliations. Model hyperparameters were optimized using a grid search for $F_2$-score and are listed in Table\textsuperscript{1}.

The individual algorithms for the ensemble were chosen by first evaluating their standalone performance on the validation set. Only algorithms available in scikit-learn, and known to perform well on natural language processing (NLP) tasks, were considered. Individual algorithms that performed well were chosen as candidates for the final ensemble model, and the best performing combination of algorithms was determined using a grid search.

**Convolutional Neural Network**

A Convolutional Neural Network is a type of deep neural network that is commonly applied to image processing tasks. Recently, however, CNNs have also been shown to be effective for various NLP problems, including text classification. The neural network architecture used in this paper (Figure\textsuperscript{1}) is based on the CNN architecture presented by Kim et al.\textsuperscript{12}. In their paper, Kim et al. demonstrate that this architecture is effective for sentence classification. The

\textsuperscript{‡}List of selectively indexed journals of interest to NLM indexers can be downloaded from the paper GitHub repository [http://github.com/indexing-initiative/selective_indexing](http://github.com/indexing-initiative/selective_indexing)
Table 1: Hyperparameters selected using a grid search. Any parameters not listed in the table were set to the scikit-learn default values.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli Naive Bayes</td>
<td>alpha=.01</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>C=2</td>
</tr>
<tr>
<td>Stochastic Gradient Descent</td>
<td>loss=modified_huber, alpha=.0001, max_iter=1000</td>
</tr>
<tr>
<td>Random Forest Classifier</td>
<td>n_estimators=100, criterion='gini'</td>
</tr>
</tbody>
</table>

architecture represents words as vectors and input text as the concatenation of word vectors. The network learns a set of convolutional filters that are convolved along the length of the input text to produce an activation map; filters learn to activate when they detect a specific type of feature (e.g. discriminative words or phrases) at some position in the text. The convolution operation is followed by a max pooling operation that keeps only the maximum activation of each filter. The result is a fixed length representation of the input text that is invariant to the position of the detected features. The final layer of the network is a task specific classification layer.

This paper presents a custom neural network architecture that uses a CNN to process text inputs. There are five inputs to the network: the article title, abstract, journal, publication year, and indexing year. The network generates a fixed length representation of each input and then concatenates them to construct the input to the hidden layer. The final classification layer uses a sigmoid activation function to generate a single output value between zero and one, which can be interpreted as the probability of an article being in-scope for MEDLINE. The model uses randomly initialized word vectors, dropout regularization, and batch normalization for the hidden and convolution layers.

The title and abstract inputs are processed separately using the same word embeddings and convolutional filter weights. Standard max pooling is used for the title, whereas dynamic max pooling is used for the abstract. Dynamic max pooling is implemented by first dividing the abstract into five equal length sections, and then standard max pooling is applied to each of these sections. The intention is to create a richer representation of the abstract by retaining some position information.

The journal is treated as a categorical input, and each journal is represented by a fixed length vector. Like the word embeddings, the journal embeddings are learned during training. The two year inputs are represented using the special encoding scheme shown in (Figure 2). The encoding is similar to one-hot encoding; however, positions for the year and preceding years are activated. The encoding is intended to capture the sequential nature of time and allow for
Figure 2: Illustration of the special encoding used for year inputs. The example shows how years between 2014 and 2018 would be encoded.

The model was trained on the full training set using binary cross entropy loss, the Adam optimizer, and mini-batch gradient descent. The $F_1$-score metric was monitored on the validation set, and training was stopped early when it stopped increasing. Model hyperparameters (Table 2) were chosen based on the literature and manual optimization for $F_1$-score.

Table 2: CNN hyperparameters.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>vocabulary size</td>
<td>400,000</td>
</tr>
<tr>
<td>word embedding size</td>
<td>300</td>
</tr>
<tr>
<td>title max words</td>
<td>64</td>
</tr>
<tr>
<td>abstract max words</td>
<td>448</td>
</tr>
<tr>
<td>number of convolution filters</td>
<td>350</td>
</tr>
<tr>
<td>convolution filter sizes</td>
<td>2, 5, 8</td>
</tr>
<tr>
<td>dynamic max pooling number of regions</td>
<td>5</td>
</tr>
<tr>
<td>activation function for classification layer</td>
<td>sigmoid</td>
</tr>
<tr>
<td>activation function for all other layers</td>
<td>relu</td>
</tr>
<tr>
<td>hidden layer size</td>
<td>3365</td>
</tr>
<tr>
<td>journal embedding size</td>
<td>50</td>
</tr>
<tr>
<td>dropout rate</td>
<td>0.5</td>
</tr>
<tr>
<td>vocabulary dropout rate</td>
<td>0.25</td>
</tr>
<tr>
<td>batch size</td>
<td>128</td>
</tr>
<tr>
<td>learning rate</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Final Combined Model

The final model combines the predictions of the ensemble of traditional machine learning algorithms and the CNN. To compute the output probability of the combined model, we assume that the component model predictions are independent and take the product of their output probabilities.

The final prediction of whether an article should be selected for indexing can only be made after the selection of a decision threshold. There is a trade-off between precision and recall: a high threshold will result in high precision but low recall, whereas a low threshold will result in low precision but high recall. The classifier developed in this paper is required to have close to 100% recall, and therefore a relatively low threshold was selected on the validation set. All the code, datasets and trained models required to reproduce the results of this paper are publicly available on GitHub at [http://github.com/indexing-initiative/selective_indexing](http://github.com/indexing-initiative/selective_indexing).

Results

This section presents a performance evaluation of the developed classifier on the 2018 test set. We first compare the performance of the combined model to the standalone performance of its component models. We then breakdown the performance of the combined model by journal topic.
Figure 3 shows precision-recall curves for the ensemble of traditional machine learning algorithms, CNN, and combined models. Figure 3a is the full plot and Figure 3b is a zoomed in plot showing recall values close to 100%. Figure 3b shows that, for intermediate values of recall, the CNN is the best performing model. In this recall range, the relatively low performance of the ensemble degrades the performance of the combined model. We are, however, interested in performance at high recall values. For the high recall values shown in Figure 3b, it can be seen that the predictions of the ensemble and the CNN model are complementary and that the combined model has the highest precision for most values of recall.

After discussion with NLM indexers, it was decided that recall of about 99.5% resulted in a good balance between classifier precision and false negative error rate. After selecting a decision threshold on the validation set, the final combined model precision and recall was measured to be 38.0% and 99.4% respectively, on the test set.

In order to breakdown model performance by journal topic, each selectively indexed journal was assigned to one of four journal groups: Chemistry, Science, Jurisprudence, or Biotech. Both the compilation of the list of journal groups and the assignment of journals to groups was done based on human judgement; the MeSH terms that are assigned to MEDLINE journals were used as guidance. Figure 4 plots precision-recall curves for the combined model by journal group and shows that the model performance varies significantly between journal groups. Specifically, the model performs better on Science and Jurisprudence articles and worse on Biotech and Chemistry articles.

Discussion

At the measured precision of 38.0%, the implemented system offers NLM indexers considerable time and cost savings by allowing them to automatically discard the 54% of articles that are very unlikely to require indexing. In 2017 there were approximately 80,000 articles processed from selectively indexed journals, and we would therefore expect the system to exclude approximately 40,000 articles from manual review each year. The measured recall value of 99.4% indicates that approximately 0.6% of in-scope articles will be missed by the system, but this is considered acceptable given the expected time-savings. It is important to realize that even human indexers may miss articles or disagree on whether an article is in-scope.

In the results section it was shown that the model performance varies considerably with article topic. The reason for this variation in performance is not explored in this paper, but the analysis highlights that the model could potentially be deployed for a subset of journals for which it is particularly effective.

One of the key challenges that was faced when developing the presented classifier was the time-variance of the dataset. There are many factors that cause time-variance and these include changes to selective indexing policy, changes to the
Figure 4: Precision-recall curves for combined model by journal group.

Figure 5: Fraction of indexed articles from selectively indexed journals against publication year. Shows the actual fraction and the fraction predicted by the CNN model.

The observed time-variance is especially problematic for the training of the CNN because it requires a large amount of training data to achieve high performance. There is insufficient data from recent years, and consequently the CNN is forced to model the time-variance of the dataset and to generalize across years if possible. The publication year and indexing year inputs were added for this purpose and were found to improve performance on the 2018 validation set. Figure 5 shows that the indexing rate predicted by the CNN closely follows the true indexing rate, and this provides further evidence that the CNN is effectively modeling the time-variance of the dataset.

Conclusion

This paper presents a machine learning based system that has been developed to assist indexers with the selection of articles for MEDLINE indexing. At the core of the system is a high recall classifier for the identification of journal articles that are in-scope for MEDLINE that combines the predictions of traditional machine learning algorithms and a Convolutional Neural Network. The system is shown to offer very significant time and cost savings by allowing
indexers to discard 54% of articles that are very unlikely to require indexing.

For future work, we plan to further explore the effect of the dataset time-variance on model performance. It is claimed that new language model based text representations require less task specific training data. It may therefore be possible to achieve better performance by training models using these new representations only on recent data. We would also like to understand why the model performance varies so much with article topic. Our motivation is that the two worst performing groups (Biotech and Chemistry) make up over two thirds of 2018 articles. Any performance improvements for these two groups will result in a significant increase in overall performance.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

We would like to thank Susan Schmidt, Melanie Huston, and Olga Printseva for their invaluable advice on selecting articles for indexing.

References


Requirements for a quality dashboard: Lessons from National Clinical Audits

Rebecca Randell, PhD¹, Natasha Alvarado, PhD¹, Lynn McVey, PhD¹, Roy A. Ruddle, PhD¹, Patrick Doherty, PhD², Chris Gale, PhD, FRCP¹, Mamas Mamas, PhD, FRCP³, Dawn Dowding, PhD, RN, FAAN⁴

¹University of Leeds, Leeds, UK; ²University of York, York, UK; ³Royal Stoke Hospital, Stoke on Trent, UK; ⁴Manchester University, Manchester, UK

Abstract

Healthcare organizations worldwide use quality dashboards to provide feedback to clinical teams and managers, in order to monitor care quality and stimulate quality improvement. However, there is limited evidence regarding the impact of quality dashboards and audit and feedback research focuses on feedback to individual clinicians, rather than to clinical and managerial teams. Consequently, we know little about what features a quality dashboard needs in order to provide benefit. We conducted 54 interviews across five healthcare organizations in the National Health Service in England, interviewing personnel at different levels of the organization, to understand how national (UK) clinical audit data are used for quality improvement and factors that support or constrain use of these data. The findings, organized around the themes of choosing performance indicators, assessing performance, identifying causes, communicating from ward to board, and data quality, have implications for the design of quality dashboards, which we have translated into a series of requirements.

Introduction

Dashboards are a type of health information technology (HIT) that use data visualization techniques to support clinicians and managers in viewing and exploring data on processes and outcomes of care¹. A distinction can be made between clinical dashboards and quality dashboards. Clinical dashboards provide feedback to individual clinicians on their performance compared to standards or targets, with the aim of informing decisions about, and thereby improving, patient care². For example, a clinical dashboard may seek to reduce inappropriate prescribing of antibiotics and therefore present data to clinicians regarding how their rate of antibiotic prescribing compares to that of their colleagues³–⁴. In contrast, quality dashboards show performance at the ward or organizational level to inform operational decision making and quality improvement efforts⁵. Ideally they will provide feedback that can be used at both ward and organizational levels⁶–⁷. It is thought that the visualizations provided by quality dashboards can lead to the identification of previously unnoticed patterns in data⁸, informing quality improvement initiatives, and more efficient and effective decision making⁹.

Healthcare organizations worldwide are increasingly using quality dashboards as a way of monitoring the quality of care they provide and as the basis for quality improvement initiatives. For example, use of quality dashboards has been reported in the US⁷, ¹⁰, Canada¹¹, the UK⁴, and the Netherlands¹². This is a trend that is likely to continue, with electronic health records providing a source of data to automatically populate quality dashboards¹³. Quality dashboards have been shown to have positive effects on performance indicators¹⁴ and use of quality dashboards has been identified amongst boards of high-performing hospitals¹⁰. However, empirical evidence regarding the impact of quality dashboards remains limited. In a systematic review of the evidence base for the use of quality and clinical dashboards, 11 research studies that had evaluated their impact on either quality or clinical outcomes were identified¹ but only one of these studies met the definition of a quality dashboard¹⁴. Consequently, we know little about what features a quality dashboard needs in order to stimulate quality improvement.

Research on audit and feedback provides a number of suggestions that are likely to have relevance when considering the design of a quality dashboard. For example, Contextual Feedback Intervention Theory (CFIT) suggests that the recipient of the feedback needs to perceive the standard or target as desirable (goal attractiveness) and achievable (goal expectancy) and perceive the feedback about the discrepancy between their performance and the standard as accurate¹⁵. CFIT contends that such feedback is more likely to change behavior if it is timely, frequent, cognitively simple, e.g. presented graphically, unambiguous, and provides concrete suggestions of how to improve performance. Hysong et al.’s model of actionable feedback suggests clinicians are more likely to respond to the feedback if it is perceived to be timely and non-punitive and if they receive feedback about their own individual performance rather than aggregated data about overall performance of the organization¹⁶. If they are able to customize how they view the data, this leads to active engagement in sensemaking, further increasing the likelihood that the feedback will be acted on. However, as the audit and feedback literature is largely concerned with feedback to individual clinicians¹⁷, this...
raises the question of how such recommendations can be adapted for quality dashboards, given that quality dashboards provide feedback to teams, whether clinical teams or managerial teams.

In the UK, the Healthcare Quality Improvement Partnership (HQIP) centrally develops and manages a programme of over 60 national clinical audits each year through the National Clinical Audit and Patient Outcomes Programme (NCAPOP). The National Health Service (NHS) standard contract requires all healthcare organizations that provide NHS services to participate in the NCAPOP national clinical audits and they are required to contribute to their funding. In addition, there are over 50 independent national clinical audits that are not part of the NCAPOP, funded either through subscription, by a charity or professional body, or by NHS England (the body responsible for commissioning NHS services in England). In comparison to Hospital Compare\textsuperscript{18} in the US, which provides an overall hospital rating made up of 57 measures, in the UK each national clinical audit focuses on a particular clinical area or condition, making them more comparable to national registries. Each national clinical audit produces an annual report that is made available to the public. National clinical audits aim to systematically measure the quality of care delivered by clinical teams and healthcare organizations and to stimulate quality improvement\textsuperscript{19}.

While there is evidence of positive impacts of national clinical audits\textsuperscript{20-22}, there is variation both within and between healthcare organizations in the extent to which they engage with national clinical audit data\textsuperscript{23, 24}. Consequently, national clinical audit data are substantially underutilized and the potential for national clinical audit data to inform quality improvement is not being realized. In response to this, we are undertaking a study to develop a quality dashboard for exploring national clinical audit data by both clinical teams and managers. As part of this, we undertook an interview study with clinicians, managers, and audit support staff to understand how national clinical audit data are currently used for quality improvement and factors that support or constrain the use of these data.

In this paper, we draw on the interview data to identify themes that have relevance for the design of quality dashboards. We first describe the methods of the interview study and then present our results. We conclude by discussing the implications of our findings for the design of quality dashboards which could be used across healthcare contexts, translating them into a series of requirements for future dashboard design.

**Methods**

**Participants**

Interviews were conducted across five acute healthcare organizations in England. Three were teaching hospitals and two were district general hospitals. To ensure our findings were generalizable beyond a single national clinical audit, interviews explored use of the Myocardial Ischaemia National Audit Project (MINAP) and the Paediatric Intensive Care Audit Network (PICANet), which vary in terms of clinical specialty, patient groups, and performance indicators. MINAP has been running continuously since 2000 and data are contributed by all hospitals in England, Wales and Northern Ireland that admit patients with acute coronary syndromes\textsuperscript{25}. PICANet was set up in 2002 and contains data from all NHS Pediatric Intensive Care Units (PICUs) in England and Wales\textsuperscript{26}. Examples of the performance indicators included in these audits are provided in Table 1. As the two district general hospitals did not have PICUs, use of PICANet data was investigated in PICUs across three of the healthcare organizations and use of MINAP data was investigated within cardiology services across all five organizations. As both MINAP and PICANet are part of the NCAPOP, to capture further variation, we also gathered information on use of the British Association of Urological Surgeons (BAUS) audits and the National Audit of Cardiac Rehabilitation (NACR), which are independent audits.

**Table 1. Example performance indicators for MINAP and PICANet.**

<table>
<thead>
<tr>
<th>MINAP</th>
<th>PICANet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance measures</strong></td>
<td><strong>Outcome measures</strong></td>
</tr>
<tr>
<td>Call (by patient/relative to emergency services) to balloon (percutaneous coronary intervention) time;</td>
<td>30 day mortality for ST-elevation myocardial infarction patients</td>
</tr>
<tr>
<td>Door (arrival in Emergency Department) to balloon time; time to angiography;</td>
<td>Complications such as bleeding, stroke</td>
</tr>
<tr>
<td>Medication on discharge;</td>
<td>Risk adjusted standardized mortality ratio;</td>
</tr>
<tr>
<td>Patient seen by a cardiologist</td>
<td>Emergency readmission within 48 hours</td>
</tr>
<tr>
<td>Accidental extubation;</td>
<td>Length of stay;</td>
</tr>
<tr>
<td></td>
<td>Admissions (planned or unplanned);</td>
</tr>
<tr>
<td></td>
<td>Refusal of admission</td>
</tr>
</tbody>
</table>
A combination of purposive and snowball sampling was used. In each site the clinical contact for the study (typically a MINAP or PICANet lead) was interviewed first. These contacts were then asked to identify others who were involved with national clinical audits, enabling us to map the networks of stakeholders through which data for particular audits were captured, accessed, and analyzed. We sought to interview participants at all levels of the organization, including clinical team members, Quality & Safety Committee members, and members of the Board of Directors. We also sought to interview those who commission healthcare services from the five sites.

Ethics approval for this study was received from the University of Leeds, School of Healthcare Research Ethics Committee (Approval no. HREC16-044).

Data collection

A semi-structured interview topic guide was developed by the research team. This was reviewed by the study Lay Advisory Group and revised based on their feedback to ensure that the interviews explored topics that matter to patients. Interviews began with discussion of the participant’s role, with a focus on their responsibilities for national clinical audits (e.g. collection, analysis, dissemination). Participants were then encouraged to draw on their own experiences and articulate how national clinical audit data were used, with an emphasis on the role of the data in informing quality improvement. They were asked about the circumstances that they felt constrained or supported use of national clinical audit data for quality improvement and what features they would consider important in a quality dashboard for exploring national clinical audit data. Example interview questions are provided in Table 2. Interviews were audio recorded and transcribed verbatim.

Table 2. Example interview questions.

<table>
<thead>
<tr>
<th>Question</th>
<th>Probes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which [MINAP/PICANet] data do you think are important for making assessments of care quality and informing quality improvement initiatives?</td>
<td>Probes: How do these measures help you to make an assessment of care quality? What are you comparing against (national average, reference standard)? Probe for differences in the perception of value/actionability of process versus outcomes data</td>
</tr>
<tr>
<td>How are the [MINAP/PICANet] data used?</td>
<td>Probes: Who accesses the data? In what format are the data accessed? What is focused on and why? Is there a process by which audit data are regularly reviewed? Do you have processes for disseminating the information to others?</td>
</tr>
<tr>
<td>Have the [MINAP/PICANet] data been used to make any changes in practice/inform quality improvement initiatives? (Explore through examples if possible)</td>
<td>Probes: How are decisions made on what changes need to be made/who was involved? Why were the changes instigated? How were the changes introduced? Were they made in a timely manner? Do you think the changes were successful?</td>
</tr>
<tr>
<td>Are there any obstacles/challenges to using [MINAP/PICANet] data?</td>
<td>Probes: How data are accessed? How data are presented? Are senior management/staff supportive and engaged in the use of audit data? If data quality, what gets in the way of data quality? If lack of timely data, how would more timely data improve things?</td>
</tr>
<tr>
<td>Do you think a quality dashboard for exploring [MINAP/PICANet] data would be useful?</td>
<td>Probes: If not, why not? What data will it need to present to be useful? How should data be presented? What features will the dashboard need to have to be useful?</td>
</tr>
</tbody>
</table>

Data analysis

The interviews were anonymized and entered into NVivo 11. Framework analysis, an approach developed for analyzing qualitative data for applied policy research, was used. Informed by the interview topic guide and reading of preliminary interviews, codes for indexing the data were identified and agreed by four members of the research team. They then indexed 5 transcripts to test the applicability of the codes and assess agreement. Where there was variation in the indexing, the codes were refined and definitions were clarified. The refined codes were applied to all
transcripts. Example codes are provided in Table 3. The indexed data were summarized in a matrix display to build up a picture of the data as a whole29. In the final stage, mapping and interpretation, the matrix was used to identify similarities and differences in participants’ responses.

Table 3. Example codes for indexing data.

<table>
<thead>
<tr>
<th>Supports or constraints on engaging with audit data</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Presentation of data</td>
</tr>
<tr>
<td>b. Timeliness of data</td>
</tr>
<tr>
<td>c. Data quality</td>
</tr>
<tr>
<td>d. Access to data</td>
</tr>
<tr>
<td>e. Dissemination and monitoring processes</td>
</tr>
<tr>
<td>f. Resources</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality dashboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Functional tasks</td>
</tr>
<tr>
<td>b. Data requirements</td>
</tr>
<tr>
<td>c. Mandated constraints</td>
</tr>
<tr>
<td>d. Look and feel requirements</td>
</tr>
<tr>
<td>e. Usability requirements</td>
</tr>
<tr>
<td>f. Performance requirements</td>
</tr>
<tr>
<td>g. Security requirements</td>
</tr>
<tr>
<td>h. Implementation requirements</td>
</tr>
<tr>
<td>i. Potential impact of a quality dashboard</td>
</tr>
</tbody>
</table>

Results

Fifty-four interviews were conducted between November 2017 and June 2018. Interviews ranged from 33 minutes to 1 hour 29 minutes, with an average (mean) length of interview of 54 minutes. Table 4 provides a summary of participants by professional group and audit. A significant number of participants, such as members of Quality & Safety Committees and Board members, used multiple audits.

Analysis of the interviews revealed five main themes that have significance for the requirements for a quality dashboard: choosing performance indicators, assessing performance, identifying causes, communicating from ward to board, and data quality. The requirements for each theme are summarized in Table 5.

Choosing performance indicators

A key issue in the design of any dashboard is the choice of performance indicators to be displayed. In discussing this with our interviewees, they talked about wanting a dashboard that was customizable, in terms of enabling them to select what performance indicators are displayed (Requirement 1). Such an approach would enable users to select performance indicators that they perceive as important signifiers of care quality. In looking at use of national clinical audit data, we found that there was limited engagement from nursing staff in part because the national clinical audits did not capture what they considered to be important markers of care quality. For example, in the PICUs, nurses were concerned about the number of central line infections, data that are not captured in PICANet. When deciding which performance indicators should be displayed on a quality dashboard, it is important that an interdisciplinary decision is made and that the selected performance indicators have relevance to all members of the multidisciplinary team.

Interviewees also talked about wanting to have access to performance indicators from other areas of the organization, a topic we return to under ‘Identifying causes’.

738
Table 4. Participants by professional group and audit.

<table>
<thead>
<tr>
<th>Professional group</th>
<th>MINAP</th>
<th>PICANet</th>
<th>BAUS</th>
<th>NACR</th>
<th>Multiple audits</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td></td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Nurse</td>
<td>6</td>
<td>3</td>
<td></td>
<td>2</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Audit support staff</td>
<td>1</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Board members</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Quality &amp; Safety staff</td>
<td></td>
<td>8</td>
<td></td>
<td>6</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Information staff</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Commissioners</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>14</strong></td>
<td><strong>12</strong></td>
<td><strong>1</strong></td>
<td><strong>2</strong></td>
<td><strong>25</strong></td>
<td><strong>54</strong></td>
</tr>
</tbody>
</table>

Assessing performance

Quality improvement depends on using the selected indicators to make comparisons, thereby providing an assessment of performance, as this interviewee describes:

*Quality improvement is about having a process in place where you are measuring those things, whether it’s patient surveys or hard data-type audit, and comparing it against either a set national standard or your peers or benchmarking [...] and that you then have a system in place where that comparison allows you to reflect on how you are performing and make the necessary adjustments to improve performance.* (Site 2, Cardiologist 4)

We identified three key ways in which clinical teams assessed their performance, which have implications for the information that should be provided in a quality dashboard and how data should be presented. One way in which performance was assessed was comparing practice against evidence-based standards that are known to improve patient outcomes (Requirement 2):

*I do reports, like monthly and yearly, to show how we’re doing. [...] it’s the nurses’ data, but I do all the data to tell them if they’re on schedule to hit the 60% target [angiogram within 72 hours for non ST-elevation myocardial infarction (NSTEMI) patients] and stuff like that. I look at all the figures for the STEMIs [ST-elevation myocardial infarction], to see if they’re hitting the 150 minute call-to-balloon and 90 minute door-to-balloon.* (Site 4, MINAP Audit Assistant)

However, such an approach only works in clinical areas where there are evidence-based standards against which to assess performance. Additionally, research on audit and feedback interventions aimed at individual clinicians emphasizes that clinicians need to see meeting these standards as achievable.

The second way in which performance is assessed is through monitoring trends over time (Requirement 3). This type of monitoring is particularly important for clinical areas where there are no evidence-based standards against which to compare service performance, as is the case in pediatric intensive care. Consequently, we found that clinical teams would use PICANet data to monitor trends over time (month-on-month):

*We use the PICANet data to produce monthly graphs that we use at our monthly clinical governance meetings. So basically, we look at re-admissions within 48 hours of discharge, deaths, accidental extubations. So, we get graphs with month-on-month numbers, so we can monitor it and see changes.* (Site 1, Pediatrician 2)

The interviewee described how monitoring PICANet data in this way led to a quality improvement initiative to reduce the number of accidental extubations; having identified a ‘spike’ in accidental extubations, additional training was introduced for nursing staff, about how to tape the tube to keep it in position and how to check the tube position on a chest x-ray.
While such assessments of performance over time were only undertaken in those sites where they had the resources to generate appropriate graphs, particularly in terms of staff skills and time, a quality dashboard should allow easy identification of such trends. Interviewees also talked about wanting to be able to choose over what time period a particular metric was displayed (Requirement 4); for example, for metrics where you may expect to see fluctuations over a year, it would be necessary to view data for several years.

The third way in which performance was assessed was through comparison with other healthcare organizations (Requirement 5). National comparisons offer opportunities to identify high-quality services in other organizations, which could act as examples of good practice from which clinical teams can learn. Making such comparisons was important to all sites, and perhaps especially to PICUs, which do not have access to evidence-based standards and thus feel the need to benchmark themselves against other units to assess their performance.

Clinical teams, in addition to comparing themselves against the national average, may also want to be able to select particular organizations to compare themselves against, organizations that they consider similar in terms of size and/or case mix (Requirement 6). One interviewee made the following comment:

"There’s PICUs that, you know, if they had no re-admission rate, I wouldn’t be surprised because it’s not a high level PICU. So, you’d have to be stratified to a PICU, an average of a PICU who is like you. So, we have cardiac patients, we have neuro patients, we have ECMO patients, so we can only really be compared with them, otherwise we’re not comparing like with like. (Site 1, Pediatrician 2)"

**Identifying causes**

Having assessed performance, a quality dashboard should then support clinical teams to identify causes for any particular performance indicators that raise concern. Interviewees described wanting to be able to ‘drill down’ into the data to understand the reasons behind the numbers, for example, to look at a particular sub-group of patients (Requirement 7). Drilling down may also go as far as looking at details of individual patients:

"It’s looking at the complications we’ve had and how we can improve them. So it’s things like the patient’s died: was it due to the procedure or was it something else? And we look at all the pathway, you know, should they have come to us, should they have come quicker? You know, we then tweak things to improve it. Or it might be that they’ve had a dissection or a blockage while they’ve been doing the procedure, and we look at, well, what equipment was used? And again, we feed back to each individual operator on their cases. So it might be Doctor so-and-so: ‘we think you should have used this piece of kit when you got into that difficulty, we’ve tried…you know, another colleague’s tried this and it worked’. So then they’ll know next time to think it. (Site 1, MINAP Audit Support Nurse)"

Interviewees at the clinical team level also talked about wanting to see data from other clinical areas, rather than data in silos, to understand how their performance interacted with and was impacted on by other areas of the organization (Requirement 8). For example, if a PICU receives a large number of emergency readmissions from the High Dependency Unit (HDU), they may also want to look at performance indicators for the HDU.

Some interviewees at the clinical team level talked about wanting a dashboard that could support simultaneous interaction by multiple users, so as to be able to drill down into the data and discuss within a meeting (Requirement 9). The interviewees felt that, through this, a quality dashboard could enhance the ability of teams to engage flexibly and in depth with national clinical audit data:

"The potential is you could use it in further discussions or using that information to probably motivate you or stimulate you into something in an audit group, so it might be that there's an aspect of care which [...] another unit might be doing really well so you might want to say: well, how are they doing that? So having that access, that would probably be something which we’d do on a larger group, that’s what I would think. (Site 1, PICU nurse)"

This may mean making the dashboard available on large touchscreens or tabletop computers, or enabling multiple users to access it via laptops and handheld devices.

**Communicating from ward to board**

For monitoring of care quality, it is important that data such as those contained within national clinical audits move from "ward to Board". However, the requirements at these different levels of the organization vary significantly. At divisional and corporate (Board and sub-committees that report to the Board, such as Quality and Safety Committees) levels, a different set of performance indicators will be of concern, those which reflect the performance of the division or organization as a whole. For example, interviewees at these levels discussed looking at performance indicators related to mortality, never events, harms, and complaints. There may be a disconnect between the performance
indicators considered at divisional and corporate levels, which are likely to reflect national priorities, and those that are considered at the clinical team to reflect care quality.

Nonetheless, our interviewees reported that there are occasions when the national clinical audits are considered at the corporate level, particularly where a clinical area is identified as an outlier; as one Board member at Site 1 said ‘Nobody wants to be an outlier.’ One way in which a quality dashboard could support the use of such data at corporate levels is by enabling easy identification of when a clinical area is an outlier within a particular audit (Requirement 10):

*What I need is flags that tell me and direct me to things that are outliers that I need to look at and then I can interrogate them [...] all of us are so pressured that we absolutely need things that are going to say: ‘problem here, here’s how you look for it’, and then you go in and look for it.* (Site 2, Quality & Safety Committee member)

One suggestion for achieving this was the use of the ‘traffic light’ or ‘RAG-rating’ system, where red, amber or green are used to indicate performance that falls below, almost falls below, or meets required standards. This was favored by several interviewees, because it is used frequently within healthcare organizations and enables users to identify potential problems at a glance. However, several interviewees also pointed out limitations with such an approach, most notably that it does not capture nuances within data, as when a standard has not been achieved for sound clinical reasons.

While it has been suggested that high performing hospitals are those where Boards spend time discussing quality issues, we found that, at divisional and corporate levels, full agendas meant there were significant constraints on time available in meetings for discussion. Where clinical team level performance indicators are considered in meetings, they will not be considered in detail. Consequently, the level of interaction and the ability to drill down that is required at the clinical team level is not needed within divisional and corporate level meetings (although will be required outside of meetings).

**Data quality**

Data quality was a topic that was frequently discussed in the interviews, with key issues being timeliness and trust in the data. At all levels, a key constraint in use of national clinical audit data for quality improvement was lack of access to timely data, with the data contained within annual reports often being perceived as too out of date to be of use. Timely data, that reflects current performance, were considered essential if they are to inform quality improvement (Requirement 11). While organizations may have timely data about their own performance, timely comparator data are also needed.

For clinical teams to engage with data for quality improvement, it is important that they have trust in the quality of the data (Requirement 12). For example, in one of the sites, they employed a dedicated PICANet audit co-ordinator who entered data into the Access database with only a day’s delay and, with support from the clinical lead, checked the accuracy and completeness of the data. Given these resources and supports, the pediatricians had confidence in the accuracy of data reports generated from the local database and used them to inform clinical governance meetings.

Trust in the quality of national comparator data is also important for clinical teams. Inconsistent or inaccurate coding was reported to have a negative impact on interviewees’ trust in MINAP data and on their ability to make meaningful comparisons with other organizations. Indeed, owing to inconsistent coding, one interviewee likened such assessments to ‘comparing apples and oranges’ (Site 3, Cardiologist).

**Discussion**

Quality dashboards are increasingly being introduced into healthcare organizations as a tool to support quality improvement, yet limited empirical research on use of quality dashboards means that we know little about what features a quality dashboard should have in order to provide the desired benefits. We have drawn on the experience of a range of professional groups who have experience of working with national clinical audit data in the NHS in England to understand what supports and constrains the use of such data for quality improvement. The findings suggest a number of requirements for a quality dashboard, which we have summarized in Table 5.

A number of the findings echo findings in the audit and feedback literature concerning the characteristics of feedback that are most likely to trigger a change in behavior. For example, the need for feedback to be timely is described in CFIT and the model of actionable feedback, and is also highlighted more generally in the literature. Similarly, credibility of the information is also noted as an important characteristic. The desire of interviewees to be able to select the performance indicators displayed can be seen as contributing to the ‘goal attractiveness’ that CFIT describes as...
influencing the likelihood of action because it enables users to select performance indicators that they perceive as important signifiers of care quality. It also enables them to select performance indicators that are consistent with established goals and priorities, another feedback characteristic that has been highlighted in the literature. Similarly, comparing performance against evidence-based standards, those that are known to lead to improved patient outcomes, may also contribute to enhancing goal attractiveness.

However, the findings also highlight some differences that have relevance when designing HIT to provide feedback to a team, rather than to an individual. For example, selection of performance indicators needs to be agreed through an interdisciplinary approach and it is necessary to ensure that they have relevance to all members of the multidisciplinary team. To support in depth engagement and discussion as a team, support for simultaneous interaction by multiple users is likely to be required and there are also implications for the hardware on which the dashboard is displayed. Literature on audit and feedback suggests that feedback should be constructed through social interaction, which is hard when it is an individual clinician interacting with a clinical dashboard but which can and should be facilitated by a quality dashboard. We also see differences between the requirements when providing feedback to an individual clinician about their performance and when providing information to managers about the performance of a clinical area. Here what becomes important is easy identification of when a clinical area is not meeting particular targets or standards.

**Table 5. Requirements for a quality dashboard.**

<table>
<thead>
<tr>
<th>Theme</th>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choosing performance indicators</td>
<td>1. Allow users to select which performance indicators are displayed</td>
</tr>
<tr>
<td>Assessing performance</td>
<td>2. Where evidence-based standards exist, make it easy to assess how performance compares to that standard</td>
</tr>
<tr>
<td></td>
<td>3. Support identification and evaluation of trends over time</td>
</tr>
<tr>
<td></td>
<td>4. Allow users to select the time period over which performance indicators are displayed</td>
</tr>
<tr>
<td></td>
<td>5. Support comparison against the national average</td>
</tr>
<tr>
<td></td>
<td>6. Allow users to select particular organizations to compare with</td>
</tr>
<tr>
<td>Identifying causes</td>
<td>7. Enable users to ‘drill down’, e.g. to look at particular sub-groups of patients</td>
</tr>
<tr>
<td></td>
<td>8. Provide access to information about other clinical areas within the organization</td>
</tr>
<tr>
<td></td>
<td>9. Support simultaneous interaction for discussion at the clinical team level</td>
</tr>
<tr>
<td>Communicating from ward to Board</td>
<td>10. Enable easy identification of when a clinical area is an outlier within a particular audit</td>
</tr>
<tr>
<td>Data quality</td>
<td>11. Provide timely data</td>
</tr>
<tr>
<td></td>
<td>12. Use sources of data that staff trust</td>
</tr>
</tbody>
</table>

We have described a number of ways in which clinical teams assessed performance: comparing performance to a standard, assessing change over time, and comparing performance to performance of other healthcare organizations. While which of these is most effective remains to be determined, we did identify that the choice of approach and
choice of organizations against which to compare will depend on the clinical area. Specifically, comparison against a standard is only likely to be effective if there is evidence to demonstrate that the standard is associated with improved patient outcomes. Existing literature on audit and feedback suggests that only one method of comparison should be used, to avoid the risk of conflicting messages, e.g. if performance is shown to be improving over time but does not meet the desired standard. However, we would argue that both comparing performance to a standard and assessing change over time are important because, depending on the organization’s starting point, the standard may take years to achieve yet it is appropriate for clinical teams to see that progress is being made.

Limitations

For healthcare organizations that are looking to develop quality dashboards, a prioritized list of requirements would be helpful, particularly if resources might be limited. While we have described how the relevance of the requirements will vary according to the user and the clinical area, the qualitative approach to data collection that we have undertaken does not allow us to provide a prioritized list. In a workshop with representatives of 22 national clinical audits, Requirements 2, 4, 7 were categorized as essential by all participants and scoring of requirements identified Requirements 1, 4, 7, and 11 as top priorities. However, the perspectives of the national clinical audits who participated in the workshop may differ from the perspectives of clinical teams and staff at divisional and corporate levels, and thus it would be beneficial to repeat the exercise with them.

A limitation of this research is that data were collected through interviews. Thus, while we have identified requirements for a quality dashboard that our interviewees perceived as important in order to support quality improvement, we do not have empirical evidence to demonstrate that a quality dashboard that meets these requirements will lead to quality improvement. In later phases of this study, we will be introducing a quality dashboard into the five organizations and evaluating its impact, allowing us to assess the importance of the requirements identified.

This research was conducted in the UK, which has an established mechanism for collecting data nationally for quality improvement purposes. However, while it might be UK focused, the findings regarding the implications for developing quality dashboards are relevant across healthcare organizations and contexts.

Conclusion

This study has explored how national clinical audit data are currently used for quality improvement and factors that support or constrain the use of these data. The findings, organized around the themes of choosing performance indicators, assessing performance, identifying causes, communicating from ward to board, and data quality, provide a number of requirements to inform the design of quality dashboards. Future research will explore the extent to which the described features support engagement with national clinical audit data and stimulates quality improvement.

Acknowledgements

This research is funded by the National Institute for Health Research (NIHR) Health Services and Delivery Research (HS&DR) Programme (project number 16/04/06). The views and opinions expressed are those of the authors and do not necessarily reflect those of the HS&DR programme, NIHR, NHS or the Department of Health.

References

A Health System’s Pilot Experience with Using Mobile Social Knowledge Networking (SKN) Technology to Enable Meaningful Use of EHR Medication Reconciliation Technology

Pavani Rangachari, PhD¹, Kevin C. Dellsperger, MD, PhD², R. Karl Rethemeyer, PhD³
¹Augusta University, Augusta, GA, USA; ² AU Health System, Augusta, GA, USA; ³ University at Albany, State University of New York, Albany, NY, USA

Abstract
In fall 2016, a two-year grant was secured from AHRQ, to pilot a mobile Social Knowledge Networking (SKN) system on Electronic Health Record (EHR) Medication Reconciliation (MedRec), to enable Augusta University (AU) Health System, to progress from “limited-use” of EHR-MedRec technology, to “meaningful-use.” The rationale is that an SKN system would enable knowledge exchange on practice issues related to EHR-MedRec, across diverse provider subgroups and settings-of-care, which, in turn, is expected to increase provider engagement, promote inter-professional learning of best-practices, and provide a foundation for practice change (e.g., Meaningful Use of EHR-MedRec technology). Over a one-year period, 50 SKN Users (physicians, nurses, and pharmacists from outpatient-and-inpatient-medicine services), participated in discussing issues-related-to-EHR-MedRec, moderated by 5 SKN Moderators (senior administrators). This paper describes the health system’s experiences with this pilot initiative; and discusses lessons learned, in regard to the potential of a mobile SKN system to enable Meaningful Use of EHR-MedRec technology.

Introduction
Over the past two decades, there has been a proliferation in the use of mobile technology to improve healthcare delivery, including a dizzying array of applications targeted toward promoting patient self-management of chronic disease, and provider-to-patient communication of treatment and lifestyle management, to improve outcomes and strengthen the patient experience¹ ². By comparison however, there is limited understanding of the potential of mobile technology to enable knowledge exchange and learning across a continuum of care providers, for promoting Meaningful Use of Electronic Health Record (EHR) technology, coordination of care, and patient safety³.

Similar to issues faced in health systems across the U.S., Augusta University Health System (AU Health), based in Augusta, Georgia, faced a scenario of low physician engagement in, and limited-use of its EHR Medication Reconciliation (MedRec) technology, which translated to high rates of medication discrepancies and low accuracy of the patient’s active medication list, during transitions of care. In fall 2016, a two-year grant was secured from the U.S. Agency for Healthcare Research and Quality (AHRQ), to pilot a mobile Social Knowledge Networking (SKN) system pertaining to EHR MedRec, to enable AU Health to progress from “limited use” of EHR MedRec technology, to “meaningful use.” An SKN system refers to an online platform for enabling moderated discussions on select topics among professionals within an enterprise, to foster professional engagement and learning on those topics. Details related to the design and implementation of the SKN system at AU Health, are provided in the section below on “Methods for Implementing SKN.”

The theoretical rationale behind an SKN system on EHR MedRec (which emanates from professional complex systems and social network theories), is that it could provide a platform for tacit knowledge exchange on practice issues related to EHR MedRec, across diverse provider subgroups and settings-of-care, to highlight adverse consequences of gaps in practice for patient safety (e.g., not using the electronic medication history function resulted in an error in recording dosage upon admission, which resulted in adverse event for the patient). This, in turn, is expected to increase physician (provider) engagement in addressing issues related to EHR MedRec; and promote inter-professional learning of best practices related to EHR MedRec (e.g., using the electronic medication history to generate the current medication list during each encounter), to provide a foundation for practice change, e.g., Meaningful Use of EHR MedRec technology ⁴ ¹⁵.

This paper describes AU Health’s experiences with novel use of mobile technology to pilot an SKN system on EHR MedRec. It also discusses lessons learned in regard to the potential of an SKN to enable inter-professional learning and Meaningful Use of EHR MedRec technology; and identifies practice implications for healthcare managers.
Problem of Interest

Healthcare delivery and payment reform efforts are increasingly focused on improving quality and safety during transitions of care, when patients are most vulnerable to medical errors. Transitions of care (from outpatient to inpatient, and back to outpatient settings), are commonly experienced by patients with chronic conditions (e.g., heart disease, diabetes, and stroke). Medication errors in particular, are common at hospital admission and discharge, and are a major contributor to adverse patient outcomes and increased spending associated with transitions of care. According to the Institute of Medicine, medication errors injure over 1.5 million people and cost billions of dollars each year. Additionally, the average hospitalized patient is known to experience at least one medication error per day. The risk of medication errors is heightened during care transitions because, clinicians, and in some cases, patients, do not have access to accurate up-to-date medication lists. This can result in the inadvertent addition, omission or duplication of medications, resulting in “unintended discrepancies” between what patients should be prescribed, and what they are actually prescribed. To help prevent medication errors and discrepancies during transitions of care, patient safety advocates have long promoted the use of Medication Reconciliation.

Medication reconciliation (MedRec) is a formal process for creating a complete and accurate list of a patient’s current medications during transitions of care. The process of MedRec comprises several key steps: (a) develop a list of the patient’s current medications; (b) develop a list of the medications to be prescribed; (c) compare the medications on the two lists; (d) make clinical decisions to update the medication list, based on the comparison; and (e) communicate the updated medication list to both the patient/family and the next providers of care. The goal of MedRec is to reduce medication errors and discrepancies during care transitions, and provide an accurate active medication list to patients and their next providers of care, to promote patient safety and quality of care.

MedRec has been a part of the Joint Commission hospital accreditation requirements since 2005, and with the introduction of the HITECH Act in 2009, it has become part of the Electronic Health Record (EHR) Meaningful Use requirements. Meaningful Use (MU) of EHR MedRec technology refers to effective use of the EHR MedRec system by providers, to complete the MedRec process (outlined above), to reduce medication discrepancies and promote medication-list accuracy, during transitions of care.

Despite the regulatory impetus towards MU of EHR MedRec however, hospital adherence has been found to lag due to low physician engagement in EHR MedRec, stemming in part, from lack of professional consensus about which physician (e.g., hospital vs. community physician) is responsible for managing a patient’s medication list, and the value of MedRec as a clinical tool for promoting patient safety. Additionally, within the hospital context, the assignment of MedRec responsibilities among provider subgroups—multiple physicians, nurses, and pharmacists, —is often unclear, leading to inefficiency and potential for error. Consistent with these findings, several recent studies have found that although hospital EHR vendors have been enhancing MedRec functionality over time, numerous hospitals still use partially paper-based processes during care transitions. In other words, there is “limited use” of EHR MedRec technology, in health systems across the U.S, as opposed to “meaningful use.”

Rationale for SKN

Based in Augusta, Georgia, Augusta University’s Health System, AU Health, is a health care network offering comprehensive primary, specialty and subspecialty care in the region. Facilities include a 478-bed AU Medical Center, more than 80 outpatient practice sites, a Critical Care Center housing a regional trauma center and a 154-bed Children’s Hospital. The health system averages approximately 21,000 inpatient discharges and 90,000 emergency room visits per year. Medicare and Medicaid together account for over 50% of the patient care revenues.

In 2016, AU Health faced challenges with the use and implementation of its EHR MedRec system (which is powered by Cerner Inc., a federally certified EHR vendor). Although MedRec was often marked as “complete” on the EHR, before patient discharge from the hospital, AU Health leadership estimated the patient’s active medication list to be inaccurate (with discrepancies between patient’s home and hospital medication lists, in regard to drugs, dosages, and frequencies), for a majority of discharged cases. Importantly, there was consensus among health system administrators, that the EHR MedRec system was not being used effectively to communicate changes in the patient’s active medication list across the provider continuum and to patients/families. This challenge was ascribed to the general reluctance of physicians to discontinue medications that they did not originally order from the active medication list, which led to not only discrepancies in the active medication list during care transitions, but also to frustrations associated with inaccurate and incomplete medication lists, among patients and providers alike, across the continuum of care.
As such, AU Health faced a scenario that was reflective of national concerns related to use and implementation of EHR MedRec technology in hospitals and health systems, i.e., low physician engagement, translating to limited-use of EHR MedRec technology. In fall 2016, a two-year grant was secured from AHRQ, to pilot-test an SKN system on EHR MedRec, for the purpose of enabling AU Health to progress from “limited use” of EHR MedRec Technology, to “meaningful use” \(^{31,32}\).

The rationale is that an SKN system would enable tacit knowledge exchange on practice issues related to EHR MedRec, across diverse provider subgroups and settings-of-care, which, in turn, is expected to increase provider engagement, promote inter-professional learning of best-practices, and provide a foundation for practice change or improvement (i.e., Meaningful Use of EHR MedRec technology) \(^{4-15}\).

Methods for Implementing SKN

The pilot implementation of the SKN system on EHR MedRec was conducted over a one-year period at AU Health, i.e., April 1, 2017-to-March 31, 2018 (Q2 2017-Q1 2018), among diverse provider subgroups and settings-of-care, i.e., physicians, nurses, and pharmacists, based in outpatient and inpatient medicine settings at AU Health. The SKN system implemented at AU Health, consisted of several components:

- **SKN Reporting Tool**: an online form that allowed participating providers (SKN Users) to report practice issues related to EHR MedRec (e.g., challenges in obtaining complete information at admission for compiling the patient’s current medication list). The form allowed users to provide a brief description of the issue and indicate the care settings and patient conditions it applies to.

- **SKN Discussion Tool (Microsoft Yammer)**: an online platform separate from the SKN Reporting Tool, to enable moderated discussions on issues related to EHR MedRec. Microsoft Yammer is an example of an enterprise SKN system. This pilot project used the basic version of Yammer, which was available to AU as part of its Office 365 package.

- **SKN Lunch-and-Learn Sessions**: A total of 5 SKN Lunch-and-Learn sessions were held over the one-year SKN period, for participants to meet and discuss lessons learned from exchanges on SKN Yammer, in person. Advance invitations to these sessions were sent to all SKN participants.

- **SKN Periodic Email Updates**: Approximately 15 periodic progress update emails were sent by the Principal Investigator (PI) to all SKN participants, over the one-year SKN period.

Additionally, there were two types of participants in the SKN system:

1. **SKN Moderators**: included a group of 5 senior administrators who played a key role in moderating discussions on issues related to EHR MedRec among participants (SKN Users) over the one-year SKN period. The 5 SKN Moderators included the Chief Medical Officer (CMO), Chief Medical Information Officer (CMIO), two hospitalist chiefs, and the Principal Investigator (PI). A key responsibility of SKN Moderators was to bring issues related to EHR MedRec reported by individual SKN Users on the SKN Reporting Tool, for discussion by all SKN Users on Yammer. As such, a majority of threads-of-discussion on Yammer began with an issue-report (problem statement) brought to Yammer by SKN Moderators. Another responsibility of SKN Moderators was to proactively initiate discussions on topics relevant to EHR MedRec (e.g., best practices identified in the literature), and bring lessons learned from the SKN system, for discussion at regular health system meetings, including the Hospital Quality Council.

2. **SKN Users**: included 50 practitioners, i.e., physicians, nurses, and pharmacists based in outpatient and inpatient medicine services at AU Health, who agreed to participate in the SKN system. Key responsibilities of SKN Users were to 1) report issues related to EHR MedRec on the SKN Reporting Tool on an ongoing basis; 2) participate in moderated discussions on SKN Yammer, over the one-year period; and 3) share lessons learned from their ongoing participation on the SKN, with colleagues in the health system. However, SKN Users were not allowed to initiate new threads-of-discussion directly on SKN Yammer. This ability was restricted to SKN Moderators, to enable coordinated discussions of issues related to EHR MedRec on the SKN system.

Participant recruitment to the SKN system, began after the project received IRB approval from Augusta University. All practitioners in three professional subgroups, i.e., physicians, nurses, and pharmacists, within five inpatient and outpatient medicine service lines at AU Health, including Cardiology, Internal Medicine, Family Medicine, Emergency Medicine, and Hospitalist service lines, were approached by the PI for recruitment using the IRB-approved informed consent process. Cardiology was the only medicine subspecialty included in the project. A total of 50 practitioners were recruited to participate as SKN Users, including 15 physicians, 15 nurses, and 20 pharmacists, from outpatient and inpatient medicine settings. All participating physicians from Cardiology, Internal
Medicine, and Family Medicine, practiced in both inpatient and outpatient settings. Only Hospitalists were exclusively inpatient practitioners.

Participants received a detailed orientation to the SKN system, prior to its launch, including steps for accessing the SKN Reporting Tool and SKN Yammer, both of which were made available to participants through separate links within the EHR at AU Health. Both tools were also accessible through the enterprise employee web portal and through mobile devices. Additionally, Yammer was downloadable as an App. Figure 1 provides screenshots of the SKN Reporting Tool and SKN Yammer.

Figure 1. Screenshots of SKN Reporting Tool and SKN Yammer.

Of the 50 practitioners who signed up to be SKN Users, 25 were active users of SKN Yammer, with 10 or more posts over the one-year SKN period. Active users of Yammer came from all professional subgroups and care settings represented on the SKN, including 8 physicians, 8 nurses, and 9 pharmacists. Additionally, 3 of the 5 SKN Moderators (including the CMO, CMIO, and PI), each posted 40 or more messages on Yammer.

Over the one-year SKN period there were a total of 485 posted communications on Yammer, divided into 62 threads-of-discussion. Of these, 45 threads had 3 or more posts; among which, 12 threads had 10 or more posts; of which 3 threads had 25 or more posts, including 1 thread with 45 posts. There were a total of 32 issues related to EHR MedRec reported on the SKN Reporting Tool, over the one-year period. Of the 45 threads with 3 or more posts, 32 began with issues reported on the SKN Reporting Tool, brought in to Yammer by SKN Moderators. As such, all issues related to EHR MedRec that were reported on the SKN Reporting Tool, were used to launch threads-of-discussion on Yammer. All threads-of-discussion on Yammer, were open and available to all SKN Users, without restriction. Yammer recorded the name, date, and time associated with each posting.

Health System Experience

Thematic analysis of the content of inter-professional knowledge exchange related to EHR MedRec on SKN Yammer over the one-year period, revealed six broad themes that were repeated across several threads-of-discussion, in the chronological order outlined below 33-35:

1) Problem Statements
2) Problem-Solving Statements (“The How-To”)
3) IT System Education (“The What”)
4) Best-Practice Assertions (“The Why”)
5) Culture Change Assertions (“The Way-To”)
6) Collective Learning (“Aha”) Moments

Figure 2 provides an example of a discussion thread (THREAD 1) on a practice issue related to EHR MedRec on Yammer, for which inter-professional discussions progressed through the full spectrum of themes from “problem statements” to “collective learning (aha) moments,” to lay a foundation for practice change (improvement), during
the one-year SKN period. In essence, this practice issue pertained to the “Importation of the External Rx History during Patient Encounters,” which pertained to the broader theme of improving communication related to the patient’s active medication list, across the continuum of providers. “External Rx History Import” refers to the electronic importation of the patient’s medication history filled at their pharmacy, by the provider (at the start of the patient encounter), by activating the Rx History button on the EHR, which would access the Surescripts system to pull the patient’s Rx history. Surescripts is an IT company that supports e-prescription, the electronic transmission of prescriptions between healthcare organizations (HCOs) and pharmacies. According to the US DHHS, in 2014, 96% of US pharmacies used the Surescripts network. At AU Health, it was determined that 90% of patients fill their prescriptions at pharmacies that participate in the Surescripts system.

Discussions related to “External Rx History Import” began with a problem statement from an Emergency Department (ED) Nurse, in regard to the challenge of obtaining the correct medication history from the patient upon arrival, to formulate the current medication list for reconciliation. In response, an outpatient nurse suggested use of the “External Rx History Import,” functionality within the EHR, for obtaining a current medication list for reconciliation. However, this suggestion was initially met with resistance. For example, the ED nurse argued that the External Rx History was not completely accurate, and therefore could not be trusted, and another provider, a Cardiologist argued that relying on the External Rx History when the patient does not have all the ‘pill bottles,’ could create errors. These disagreements were met with problem-solving statements from other providers. For example, a hospitalist physician argued that even if an electronic list of medications cannot be fully trusted, it could provide a starting point for discussing with patients what they are actually taking. Such statements were then followed by IT system education, e.g., clarifications related to the External Rx History functionality on the EHR by the CMIO/SKN Moderator, which was followed by best-practice assertions from pharmacists, i.e., articulations of the benefits of adhering to and consequences of not adhering to the best practice of External Rx History Import. For example, a pharmacist argued that importing “External Rx History,” preferably as close to admission as possible, is essential, since it includes 90% plus of what actually occurred in the community. While there may be some gaps, it would at least allow providers to know what the patient has been prescribed and help them question why they are not taking certain medications. The CMO/SKN Moderator then helped to reinforce these best-practice assertions, by providing examples of recent cases where dosing errors could have been averted, if the External Rx History option had been used. Another pharmacist then put forth a culture change assertion, urging participants to not allow ‘perfection to become the enemy of the good,’ arguing that, when prescribers do not act to reconcile a medication owing to insufficient information, they are making a decision to ‘do nothing.’ The CMO/SKN Moderator helped to reinforce this message by arguing that ED patients and hospital-to-hospital transfers have little ability to provide medication history, making the importation of External Rx History, all the more important. These discussions culminated in a collective learning (aha) moment summarized by the PI/SKN Moderator, in that, “incremental efforts to improve the accuracy of the current medication list, by leveraging the External Rx History Import function, could go a long way in reducing medication discrepancies during care transitions.” The second SKN Lunch-and-Learn session (held in July 2017), provided an opportunity for SKN Moderators to meet face-to-face with several SKN Users. The essential purpose of this session was to summarize lessons learned from Yammer on the topic of External Rx History Import; and reinforce the collective learning (aha) moments on this topic.
Figure 2. Dynamics of Inter-Professional Knowledge Exchange on External Rx History Import (THREAD 1).
The aforementioned inter-professional learning dynamics in turn, coincided with a distinct improvement trend in the Aggregate Proportion of External Rx History Import among all Congestive Heart Failure (CHF) patient encounters in the 5 participating medicine service lines at AU Health, during and beyond the SKN period. This measure may be regarded as a “measure of Meaningful Use (MU) of EHR MedRec technology” (or a measure of best practice in EHR MedRec), because a higher proportion, reflects better use of the EHR MedRec system by providers, to obtain a current medication list, for reconciling with new prescriptions, to reduce medication discrepancies and improve accuracy of the active medication list, during transitions of care.

Figure 3 depicts the trend in this measure over a 6-quarter period (Q1 2017-Q2 2018). The vertical dotted line on the graph represents the point-in-time of the second SKN Lunch-and-Learn session, which was conducted (at the start of Q3 2017), after several weeks of inter-professional discussion on SKN Yammer. Essentially, this session served as a culminating point for collective learning on the topic of External Rx History Import. As indicated in the figure, substantial improvement (increase) of over 140% was noted in the measure, during and beyond the SKN period, including a discernible increase from Q2 to Q3 2017 among the SKN participating medicine service lines. On the other hand, the comparative data for the Neurosurgery service line at AU Health (which did not participate in the SKN system), shows a relatively stable trend in the same measure, with the aggregate proportion of “External Rx History Import” for this service line, remaining at an average of ~45%, throughout the same timeframe. These dynamics in inter-professional learning and practice improvement suggest that the distinct improvement trend noted in this measure of MU of EHR MedRec technology among the 5 participating medicine service lines, may have emanated from the inter-professional learning associated with this topic, on the SKN system.

Lessons Learned and Practice Implications

The health system’s pilot experiences suggest that a mobile technology-enabled SKN system could be a valuable tool for engaging providers to share knowledge on practice issues and promote inter-professional learning of best-practices, to create a foundation for practice change/improvement, i.e., EHR Meaningful Use. The experience also suggests that the SKN system enabled progress towards MU of EHR MedRec, by addressing EHR implementation-challenges in the correct sequence, i.e., by facilitating collective learning of the value of best-practices in EHR MedRec and a shared understanding of the system-level EHR MedRec workflow among providers (i.e., the “big picture”), before IT-training of providers to address socio-technical challenges of EHR implementation (e.g., differing system-views across provider subgroups). Results also suggest that an SKN system may be a particularly valuable tool for institutions where practice change (e.g., MU of EHR technology) needs to occur with existing resources, through workflow changes; as opposed to new resource outlays, e.g., dedicated pharmacy techs in each and every unit within the health system. Importantly, the experience suggests that the health IT components of the SKN system, e.g., SKN Yammer and the non-health-IT components, e.g., SKN Lunch-and-Learn sessions, served to not only complement, but also reinforce each other synergistically, to enable inter-professional learning, and provide a foundation for practice change (MU of EHR MedRec technology).
From a broader perspective, the experiences with this SKN pilot initiative at AU Health, help to identify the following evidence-based management strategies for the creation of “learning health systems,” for enabling successful change implementation in HCOs.

- At the start of the effort, create a knowledge sharing and learning mechanism (e.g., SKN) among a critical mass of providers, expected to implement practice changes. The mechanism should facilitate tacit knowledge exchange on issues experienced with current practices, as well as possible resolutions to those issues.
- Conduct proactive, periodic communications (from senior leadership) on benefits of adhering to best practices and consequences of not adhering to them.
- Create shared understanding of the value of best practices (i.e., the link between practices and outcomes) or the “why,” to gain provider engagement in practice change, before embarking on IT-training of providers to address sociotechnical (health IT) challenges.
- Develop capacity to collect, analyze, and disseminate data on best-practice measures among providers to promote a scientific (research-based) approach to learning and improvement.
- Enable champions for change to emerge from among providers, to voice the need for culture change for enabling successful best-practice implementation; and reinforce these messages with proactive, periodic communication from senior leadership.
- Create a “learning health system,” by synthesizing lessons learned to facilitate “collective learning (aha) moments” across provider subgroups and care settings; and encourage providers to spread the learning within the broader institution.

The aforementioned set of strategies, represents a significant contribution to healthcare management practice, particularly in the context of a growing body of literature that has argued for health IT initiatives to be evaluated using a “health system” lens, i.e., from a broader perspective of achieving health system goals. In other words, hospital IT initiatives should be able to be integrated into existing health system functions and complement health system goals to avoid being perceived as silo or stand-alone solutions. This paper provides a meaningful example of a pilot hospital IT initiative that was integrated into a broader health system goal of developing a “learning health system,” to enable Meaningful Use of EHR MedRec technology. To this effect, this paper is also directly aligned with the AMIA 2019 Annual Meeting Theme: “Informatics: From Data to Knowledge to Action.”

Conclusion

The health system’s experience with this pilot initiative suggests that an SKN system could be a valuable tool in enabling inter-professional learning to promote complex practice change (e.g., MU of EHR technology). Future research could help to evaluate the generalizability of experiences in this case, through large-scale controlled experiments to investigate causal relationships between SKN use and Meaningful Use of EHR technology, across a wide variety of HCOs. These types of research endeavors in turn, would help to generate a systematic evidence-base of strategies for promoting Meaningful Use of EHR technology, which, in turn, could be used to prompt federal EHR vendors to incorporate SKN features into EHR systems.

References

Considerations for Improving the Portability of Electronic Health Record-Based Phenotype Algorithms

Luke V. Rasmussen, MS1, Pascal S. Brandt, MSc2, Guoqian Jiang, MD PhD3, Richard C. Kiefer3, Jennifer A. Pacheco, MS1, Prakash Adekkannattu, PhD4, Jessica S. Ancker, MPH PhD4, Fei Wang, PhD4, Zhenxing Xu, PhD4, Jyotiushan Pathak, PhD4, Yuan Luo, PhD4
1Northwestern University, Chicago, IL; 2University of Washington, Seattle, WA; 3Mayo Clinic, Rochester, MN; 4Weill Cornell Medicine, New York, NY

Abstract
With the increased adoption of electronic health records, data collected for routine clinical care is used for health outcomes and population sciences research, including the identification of phenotypes. In recent years, research networks, such as eMERGE, OHDSI and PCORnet, have been able to increase statistical power and population diversity by combining patient cohorts. These networks share phenotype algorithms that are executed at each participating site. Here we observe experiences with phenotype algorithm portability across seven research networks and propose a generalizable framework for phenotype algorithm portability. Several strategies exist to increase the portability of phenotype algorithms, reducing the implementation effort needed by each site. These include using a common data model, standardized representation of the phenotype algorithm logic, and technical solutions to facilitate federated execution of queries. Portability is achieved by tradeoffs across three domains: Data, Authoring and Implementation, and multiple approaches were observed in representing portable phenotype algorithms. Our proposed framework will help guide future research in operationalizing phenotype algorithm portability at scale.

Introduction and Background
As the proliferation of electronic health records (EHRs) has increased the secondary use of healthcare data, it has also catalyzed the establishment of collaborative research networks where multiple institutions are working on large studies to increase their sample size and diversity of the study population. This has been demonstrated via multi-site prospective clinical trials and large-scale observational studies, where participant eligibility criteria queries - which, typically take the form of phenotype algorithms (i.e., structured selection criteria designed to produce research-quality phenotypes)1,2 - are run at each participating institution to identify the study cohort. More recently, large collaborative networks such as the Health Care Systems Research Network (HCSRN, formerly the HMORN)3, Mini-Sentinel4, the electronic Medical Records and Genomics (eMERGE) Network5, the National Patient-Centered Clinical Research Network (PCORNet)6, the Accrual to Clinical Trials (ACT) Network7, the Observational Health Data Sciences and Informatics (OHDSI) program8, and the Electronic Health Records for Clinical Research (EHR4CR) European project9 have shown different models for how a single phenotype algorithm may be run across heterogeneous EHRs and database systems. A desired state from many of these networks is to have “computable phenotype algorithms” that are portable, and can be executed in a high-throughput manner10.

In our observations, across both clinical trials and research networks, two primary models have emerged for the implementation of phenotype algorithms at multiple sites: 1) sharing a human-readable narrative of the algorithm that is re-implemented at each site (e.g., clinical trials, eMERGE), and 2) using a pre-coordinated common data model (CDM) which allows sharing of machine interpretable executable code (e.g., OHDSI, PCORnet, HCSRN). The first approach often reduces the burden on the algorithm authoring institution, as they may describe the phenotype algorithm logic and required medical vocabulary codes in a more generic manner or in a manner that meets their local institution-specific settings (e.g., local, non-standardized laboratory measurement codes). This algorithm logic can then be disseminated widely via public repositories - the eMERGE Network uses the Phenotype KnowledgeBase (PheKB)11, and clinical trial criteria are made available on ClinicalTrials.gov12. Sites implementing the phenotype algorithm then interpret and convert the written description to an executable format that can run against their local data, which often requires iterative communication with the original author(s) to ensure accuracy. Local validation is an important component to ensure the implementation is correct, and that the phenotype algorithm implementation is sufficiently sensitive and/or specific (depending on the specific study needs) when implemented across multiple sites13. To this end, phenotype algorithms may be easily shared as a narrative document, but this limits scalability as each algorithm must be transformed from a narrative to local queries. This has been shown to be a cumbersome, resource-intensive and error-prone process14.
The second approach, the use of a CDM, uses a pre-coordinated structure for the data such that the phenotype algorithm – represented as a database query – may be directly shared and executed at different sites using the same CDM. Several CDMs have been defined, including Informatics for Integrating Biology and the Bedside (i2b2)\textsuperscript{13}, PCORNet\textsuperscript{6}, Mini-Sentinel CDM\textsuperscript{16}, OHDSI’s Observational Medical Outcomes Partnership (OMOP)\textsuperscript{17}, the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM)\textsuperscript{18}, and the HCSRN’s Virtual Data Warehouse (VDW)\textsuperscript{19, 20}. CDMs by their nature are optimized for a specific set of uses and data sources (e.g., clinical vs. research, claims data vs. EHR data)\textsuperscript{21}. Furthermore, no CDM is “universally common,” meaning that translation and reimplementation are required when crossing CDM boundaries. A more recent project led by the Food and Drug Administration, “Harmonization of Various Common Data Models and Open Standards For Evidence Generation”\textsuperscript{22}, is developing mappings from several of these CDMs to the BRIDG\textsuperscript{23} domain analysis model to address this need. In addition, local adjustment of the phenotype algorithm is still sometimes needed to account for differences that may not be addressed by a CDM, such as different numeric ranges for abnormal lab results across sites. Similar to the first approach, sensitivity and specificity still need to be assessed to ensure the phenotype algorithm is complete (e.g., includes all relevant vocabulary codes). Because of this, portability of phenotype algorithms is not solved solely by a CDM, although a CDM provides many advantages\textsuperscript{10}.

The Phenotype Execution and Modeling Architecture (PhEMA) project has proposed an architectural model\textsuperscript{24} that supports the authoring of standards-based EHR-based phenotype algorithms, and their subsequent execution across multiple CDMs. This is done by the use of a canonical, standard representation of the phenotype algorithm logic and value sets (collections of standard medical vocabulary codes). PhEMA then automates the translation steps to convert these pieces to computable phenotype algorithms for individual CDMs. Most recently, PhEMA was used to execute a benign prostatic hyperplasia phenotype algorithm across multiple institutions’ data warehouses and i2b2 instances\textsuperscript{25}, demonstrating a basis for a single representation of a phenotype algorithm that can span multiple data models. In addition to crossing data model boundaries, we recognize that many other facets of portability exist. In this work, we build on the reported experiences of collaborative research networks, as well as our experiences in implementing PhEMA across three academic medical centers, by further exploring the considerations needed for cross-site portability of phenotype algorithms.

**Methods**

We identified large research networks published in the health and biomedical informatics literature that have focused on building or leveraging infrastructure to support cross-site execution of EHR-based phenotype algorithms. Identification of the networks was done from a cursory search of PubMed for consortia describing EHR-based phenotyping, supplemented by the authors’ collective experience. Within this paper, we are broadly considering an “EHR-based phenotype” as a definition of a disease or trait that is ascertained primarily from data stored in an EHR. We limited consideration to only those networks describing a distributed model, in which participating institutions retain access to their respective data, and those that have an established infrastructure to demonstrate portable phenotype algorithm execution. In total, we selected a purposive sample of seven collaborative networks: ACT Network, EHR4CR, eMERGE, HSCRN, Mini-Sentinel, OHDSI, and PCORnet. We reviewed the literature published by the respective networks, as well as documentation available online regarding CDMs and technical architecture. This was supplemented by the authors’ experience from participation in some (but not all) of the networks, and personal communication with members in other networks. From this, we created maps showing the elements used to compose a phenotype algorithm, the methods by which phenotype algorithms are shared across sites, and how they are executed at each site. Multiple approaches may be used within a single network; however, as our objective was to evaluate the distinct approaches, we did not exhaustively document all possible permutations across all networks. These models were consolidated into a final framework to capture and model the considerations needed for portable EHR-based phenotype algorithms.

**Results**

From review and consolidation of the respective research network models, we developed a framework for phenotype algorithm portability that has three main components: 1) a phenotype algorithm workflow model, 2) a portability tradeoff model, and 3) a portability representation model.

**Phenotype Algorithm Workflow Model**

The phenotype algorithm workflow model is comprised of 8 steps across 3 broad domains that we identified as being involved within the phenotype modeling and execution process, and that have potential impact on portability (Table 1). As described, these steps identify the core work needed to establish a portable phenotype algorithm, and are part...
of an iterative process.

Table 1. Phenotype algorithm workflow model

<table>
<thead>
<tr>
<th>Domain</th>
<th>Step</th>
<th>Description</th>
<th>Potential Challenges to Portability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Data Collection</td>
<td>The processes by which data is collected within the source EHR, and its intended purpose.</td>
<td>Only data that is collected can be used for electronic phenotyping.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>How data is collected at a local institution (vocabulary used, frequency of collection, etc.) determines how that institution authors a phenotype algorithm.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Modality of data collection (e.g., structured, narrative text, images) can affect how and if the data used in executing a phenotype algorithm.</td>
</tr>
<tr>
<td>Data</td>
<td>Preparation</td>
<td>Extract-Transform-Load (ETL) processes through which data is consolidated into an integrated data repository (IDR).</td>
<td>The need to transform the shape of the data from an IDR data into a common data model (CDM).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Effort to convert data from one modality to another (e.g., natural language processing to obtain structured results).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mapping of local terms to a standard vocabulary term (national standard or prescribed by CDM), and potential lossy mappings or semantic drift.</td>
</tr>
<tr>
<td>Authoring</td>
<td>Define Value Sets</td>
<td>Identifying the medical terms that are used to represent data elements within the phenotype algorithm logic.</td>
<td>Not all terminologies/vocabularies are fully implemented at all institutions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Value sets may list all codes, or may list codes at the top level of a hierarchy that need to be expanded.</td>
</tr>
<tr>
<td></td>
<td>Define Logic</td>
<td>Create a representation of the required data elements, and how the elements are related by different operators (e.g., Boolean, temporal) to create a phenotype algorithm.</td>
<td>The modality of the logic representation (narrative, intermediate representation, programming language), and what system(s) may understand it.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Strictness of the logic, considering local instead of broader data availability.</td>
</tr>
<tr>
<td>Implementation</td>
<td>Distribution</td>
<td>The mechanism by which a phenotype algorithm is transmitted from the author to an implementing site.</td>
<td>Automated vs. manual approach.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Policies that require human review and approval before execution.</td>
</tr>
<tr>
<td></td>
<td>Translation</td>
<td>How the phenotype algorithm is converted into an executable representation that may be directly applied to the institutional data model.</td>
<td>Automated vs. manual approach.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Technology-specific customizations (e.g., database schema names, table names).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Information loss when elements of a data model do not have a direct translation or differ in granularity.</td>
</tr>
<tr>
<td></td>
<td>Execution</td>
<td>The computation process by which the executable representation is applied to an institutional data warehouse, and results are retrieved.</td>
<td>Syntax errors that require human intervention and correction.</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>A formal or informal comparison of the execution results against a reference standard.</td>
<td>Lack of detailed information concerning the inclusion and exclusion implications across multiple phenotype algorithm steps.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lack of access to source data to evaluate results.</td>
</tr>
</tbody>
</table>
Data Collection - A reality for any research network is that if a particular data element is not collected as part of an institution’s care process, it will never be available to a phenotype algorithm. Similarly, if a data element is available within a source system but is not made available to a site’s integrated data repository (IDR) in a way that it can be queried, a phenotype algorithm will not be able to include it. This may be observed with data collected as clinical narratives or images, as opposed to structured data. We note that even with a common data model that is capable of representing all medical concepts, these potential barriers would keep the data not represented in the CDM from being provided. This remains a consistent barrier across all networks.

Data Preparation - We observed that many networks adopted a CDM to aid with portability. This requires those involved with the Data Preparation step to spend additional time “data wrangling” (i.e., cleaning, mapping and translating their source EHR data into the CDM). This reduces the downstream effort on both authors and implementers. There are variations in the pre-coordination of vocabularies within the CDMs, however. For example, within i2b2 (used in the ACT network), each site may have a local ontology specified for their data elements. This allows for custom data elements (e.g., status within a biobank or clinical trial), as well as individual site selection of the vocabulary to use (e.g., NDC codes for medications instead of RxNorm codes). Other networks, such as OHDSI, provide stronger guidance on the vocabularies to use, potentially requiring additional time and effort in the data mapping process. They still allow flexibility by preserving the source system’s native vocabulary code, and allow phenotype algorithms to be defined using those codes. Some CDMs account for processing steps to convert data between modalities, such as the NOTE_NLP table within OMOP to store structured results from natural language processing (NLP). This requires additional effort to perform NLP on clinical text, but produces structured data that is more amenable to processing. Depending on the amount of effort needed, not all sites adopting a CDM may perform this pre-processing, or may do it on an ad hoc basis as new categories of data are needed for a study (e.g., processing just echocardiogram reports). On the other hand, it was pointed out that existing NOTE_NLP table lacks flexibility to store relations between medical concepts.

Defining Value Sets - In networks where a pre-coordinated set of vocabularies have been implemented (as described in Data Preparation), time and effort are reduced within the authoring process as the phenotype algorithm author can write the phenotype algorithm for just those vocabularies. For example, within OMOP, diagnoses may be represented entirely using SNOMED CT. Within networks such as eMERGE where sites re-implement the phenotype algorithm locally, additional effort is needed by the author or implementation site to expand a value set. For example, an existing phenotype algorithm developed using just ICD-9-CM codes only would need to be updated for an implementer wishing to include ICD-10-CM codes (ICD-10 codes were used in the US only after October 2015). We note that the reduced authoring time in networks having a pre-coordinated vocabulary does not necessarily guarantee portability across research network boundaries. For example, an HSCRN query representing a medication value set with NDC codes would require mapping to RxNorm to support execution in ACT.

Defining Logic - Research networks use three general approaches to make their phenotype algorithms portable. The first is the use of a specific programming language to represent the query (e.g., SAS within HSCRN, R and SQL within OHDSI). A second approach is the use of a system-specific intermediary representation for the logic, agnostic to underlying implementation. For example, the ATLAS JSON format within OHDSI, the XML query definition in i2b2, and the Eligibility Criteria Language for Clinical Trial Investigation and Construction (ECLECTIC) within EHR4CR. The third approach is the development of a written narrative and/or visual flowchart to describe the phenotype logic (e.g., phenotype logic posted to PheKB for eMERGE). This last format is based on a computable implementation at the authoring site, which may be done against any data model and using any technology the authoring site chooses. The description of the phenotype algorithm may be easier for the author to provide as it can convey the intent of the logic without prescribing specific tables or fields to be used, but requires each implementing site to read, evaluate and map the intent of the logic to their local institutional data model.

Distribution – Phenotype algorithms were distributed manually or automatically, and the choice of distribution method often depends on how the logic was represented as well as constraints within the research network. Within the eMERGE network, the author will manually publish a phenotype algorithm to PheKB, and inform implementing sites when it is ready. Similarly within OHDSI, phenotype algorithms may be tracked within a central GitHub repository, where they can be reviewed and manually imported into a local OHDSI instance. The HSCRN and PCORnet have infrastructure to electronically share the phenotype algorithm to implementing sites, where it may be reviewed from a worklist before execution. Federated architectures, such as EHR4CR’s Orchestrator, ACT’s use of the Shared Health Research Information Network (SHRINE) on top of i2b2, or the ARACHNE system within OHDSI, allow phenotype
algorithms to be automatically distributed. Within this step, automation can reduce burden on implementers, and also reduce burden on authors if a technical solution exists to link authoring to distribution.

**Translation** - Varying levels of transformation are needed across the networks, driven by technical capabilities. Within HSCRN (where SAS files were directly shared) and within OHDSI (where SQL and R code are shared), some manual translation may be needed to configure the code to run locally. As an example, SQL code for OHDSI may require specifying a local database schema name before it can be run. Within OHDSI where the JSON representation of a phenotype algorithm is shared from the ATLAS system, the translation is done automatically. Similarly, within ACT, SHRINE performs real-time mapping of ontology terms in the XML message to the local i2b2 instance, and i2b2 itself manages translation of XML to local SQL queries.

**Execution** - This step is consistent across all research networks, as all translations have been completed and the phenotype algorithm is now in an executable format for the local data repository. During the execution phase, considerations for portability are driven largely by the results of the execution. A phenotype algorithm may be appropriately translated to SQL code, but requires a significant amount of time or resource to run. This may be to the point where the SQL code is unable to complete before a system-level timeout causes it to halt; thus, it is not truly executable and is no longer viewed as a portable implementation. Likewise, a phenotype algorithm may have execution paths that are entirely dependent on a category of data that is not present due to considerations from steps in the Data domain. Although a syntactically valid definition of the phenotype algorithm exists, it is also not viewed as portable given that it returns no records when some records are expected to be found.

**Validation** – This step is carried out in multiple ways across the different research networks, often depending on the technical approach used to perform actual execution. Usually the first pass of the execution is a simple cohort count (“how many cases and controls with X?”), which may be supplemented with a demographic breakdown. An initial face-validity check of the result can be used to ascertain if the results seem feasible, given expected prevalence of a condition. To perform more in-depth validation, or to troubleshoot why a phenotype algorithm is returning an empty cohort, access to the underlying data model is typically needed. As this is a manual process, it requires additional coordination and time across implementation sites by the author site to conduct chart reviews. Automation of this type of analysis can be done in the Execution step, although this increases overall execution time.

**Portability Tradeoff Model**

The second component of our proposed framework is a portability tradeoff model that represents the interlinked nature of the three domains in the workflow model. Across the seven reviewed research networks, we observed that different weights are given to Data, Authoring and Implementation. Within the portability tradeoff model, we consider a higher weight in a domain as having an increased overall cost (although we note that a high cost is not inherently “good” or “bad”). These costs are in turn linked to the “iron triangle” of project management: time, features and resource. This results in what we have termed the “nested iron triangle of phenotype algorithm portability” (Figure 1). For example, OHDSI and PCORnet put additional weight on the Data domain, such that more up-front work is needed at each institution to adhere to the CDM. This reduces the time spent on Authoring, as the author only needs to consider a single data model and a limited set of vocabularies. Similarly, less effort is then needed at each site for Implementation, as less localization is needed during the Translation step. As a second example, sharing phenotype algorithms as narratives (employed within the eMERGE network) allows flexibility in the data model, often leveraging existing data warehousing infrastructure (decreased weight on Data). However, additional time may be required during Authoring, if the author is considering portability and thoroughly investigating comprehensive value sets. Furthermore, more weight is given to steps in the Implementation domain, as each site must put more effort into the Translation step to arrive at an executable definition.

![Figure 1. Portability tradeoff model: the “nested iron triangle of phenotype algorithm portability”.](image-url)
**Portability Representation Model**

During our review, we observed how the seven research networks approached the representation of the phenotype algorithms, and the impact this representation had on portability across steps in the Authoring and Implementation domains. This is defined in its own model given the additional considerations needed. Four general approaches were identified (Figure 2), with their own benefits and limitations.

**Unchanged** – This approach is the ideal state of portable representation, as a single definition (the originally developed phenotype algorithm) may be distributed for use at the implementing sites. Technically this approach is currently available in our reviewed networks, with the exception of the non-computable narrative phenotype algorithms in the eMERGE network.

**Criteria Relaxation** - This approach allows making a phenotype algorithm more general by reducing or removing constraints, while monitoring and balancing sensitivity and specificity for the intended use case. This results in a single phenotype algorithm that is distributed, and only includes the logic known to be runnable at each implementing site. We observed this as being universally available across all of the research networks from a technical standpoint. All reviewed systems across the research networks allow criteria to be removed entirely (if overly-exclusive), or constraints broadened (if insufficiently inclusive). This requires creating a revision of the phenotype algorithm with the modified logic and/or value sets. This may result in a phenotype algorithm with overall lower sensitivity and higher specificity across multiple sites, but allows the phenotype algorithm to remain a singular entity that eases distribution, as the same phenotype algorithm is shared universally.

**Multi-Label Specification** - This approach involves implementing a single version of the phenotype algorithm, but allowing different labels to be applied to the cohort, depending on quality of evidence (e.g., “gold standard” and “silver standard”). Unlike criteria relaxation, this allows the single phenotype algorithm to be run at each site and produce results, but may increase overall complexity of the phenotype algorithm depending on the number of logic paths. This is not directly supported outside of narrative descriptions (eMERGE phenotype algorithms) and executable implementations (e.g., HSCRN SAS, OHDSI SQL and R). In networks that provide an intermediary platform to author phenotype algorithms (ACT’s i2b2 and OHDSI’s ATLAS), no solution was observed. Instead, authors must create multiple versions of a phenotype algorithm - one for each label. This causes duplication of logic, which must now be maintained across multiple phenotype algorithms, and these platforms do not provide a way to group the disparate phenotype algorithms together to show they are related.

**Site-Specific Customizations** - In this approach, the implementing site changes the logic or translates executable code of the phenotype algorithm so that it will run and produce acceptable results. This may be done at the discretion of the implementing site, or done in conjunction with the original author to assess tradeoffs. The result is multiple versions of the phenotype algorithm, which are (at best) loosely linked together. This approach is also universally available across all of the research networks. We note that while it provides flexibility to implementers (at the tradeoff of additional work), it severs provenance to the original phenotype algorithm. This makes subsequent changes by the original author difficult or impossible to incorporate automatically, as well as changes from the implementers that may be useful to the author and other implementers.

*Figure 2. Portability representation model: Four approaches to represent computable phenotype algorithms that have been executed across multiple sites.*

---

760
Discussion
Here we have presented a framework of phenotype algorithm portability based on the reported experiences of seven research networks. It is comprised of three interlinked models, which together provide researchers and system evaluators with a comprehensive framework against which to consider barriers and solutions to the portability of EHR-based phenotype algorithms. The framework demonstrates that the concept of phenotype algorithm portability is multi-faceted, as shown by the number of steps and potential impact points, and by the observation that the approaches within each model are not necessarily mutually exclusive within a single network. For example, within the portability tradeoff model, an institution adopting a CDM such as PCORnet may put less up-front investment in the Data domain (reducing it compared to other sites), which then would increase Implementation resource needs for that site. Similarly, within the portability representation model, multiple approaches may be observed within the same network. Within OHDSI, an ATLAS-authored phenotype algorithm may be shared across sites as unchanged, but a SQL script written for a specific database vendor may require site-specific customization to port the syntax to another vendor system. While in our framework and respective models we consider automation and computability, we note that the concept of portability is not dependent on them. For example, several of the research networks that we evaluated require the manual review of a phenotype algorithm before it could be executed (driven by that network’s operating policies). We recognize that this mandated manual review step impacts how scalable that portability solution is to a large number of phenotype algorithms.

Our proposed framework is supported by previous studies reported in the literature, but to our knowledge, no previous study has considered all aspects of portability. Our framework supports the ongoing need for localization of phenotype algorithms, which was also identified within a previously reported desiderata33. We believe the need for localization will remain for the foreseeable future, given the number of disparate CDMs that exist, and the complexities of data collection and representation within the EHR34. As CDMs continue to evolve, and as more work is done to harmonize CDMs35, that need will lessen. However, previous studies have demonstrated the impact of data heterogeneity on phenotype algorithms36, 37, which we believe will be an ongoing issue. Recognizing this, we propose adjusting the view of a phenotype algorithm as a single, immutable object. This view is currently reflected in practice, as demonstrated by three of the four approaches in our portability representation model (criteria relaxation, multi-label specification, and site-specific customization). No system to date has fully addressed how to represent this in a computable way, although we note systems like PheKB provide a communication hub to centralize storing multiple site implementations for a phenotype algorithm.

Instead, as more phenotype algorithms are made computable, they should incorporate localization considerations centrally, regardless of the approach needed. For example, criteria relaxation allows for a central computable definition, as it involves changing a single definition. The inclusion of version tracking (an approach used by PCORnet’s ADAPTABLE trial38 and OHDSI39) would also allow tracking changes to the phenotype algorithm as criteria are relaxed, and allow documentation of the tradeoff decisions considered over time. The resulting phenotype algorithm would then be a single repository with the full history available. For multi-label specification, a computable phenotype algorithm should provide a means by which the label may be assigned during authoring, and automatically tracked during implementation. This will require adjusting tools and specifications used in the authoring and execution processes. We note that within the PhEMA project, work is ongoing to develop a representation that meets these requirements. The final approach, site-specific customization, may benefit from lessons learned in version control systems for source code management. Based on a single, original phenotype algorithm, implementers may make a copy (“branch”) which retains a link back to the original. The implementer may propose changes (via “pull request”) or the author may review localized copies of the phenotype algorithm to determine if changes could be incorporated into the original definition.

With particular respect to controlled medical vocabularies, a previous review noted the overlap in available terms across vocabularies, although no clear rationale was ascertained as to why different vocabularies (with incomplete mappings observed in some instances) were used in a phenotype algorithm14. We posit, based on our experiences, that this is in part due to the availability of data at a local institution (the Data phase of our proposed model). That is, the rules developed or features learned in the creation of a phenotype algorithm are driven by the data collected and mapped to an IDR or CDM. While some phenotype algorithm authors do consider the most comprehensive value set definition up-front, and CDMs such as OMOP use a prescribed set of vocabularies to obviate the need for heterogeneous vocabulary mapping, they are not singular solutions. This further highlights the importance of
implementing and evaluating phenotype algorithms at multiple institutions, which will naturally uncover gaps in the original definition. Portability, in turn, becomes a critical consideration.

Although our review primarily considered rules-based phenotype algorithms, similar parallels can be drawn with respect to machine learning (ML)-based data-driven algorithms. ML-based approaches have demonstrated potential to achieve high-throughput phenotyping by learning relevant features of the EHR data (including features derived from narrative text), thereby reducing the time to establish a phenotype algorithm as opposed to manually curated, expert-driven rules-based phenotype algorithms at each institution. Additional considerations for portability within the phenotype algorithm workflow model would include an expansion of our proposed Translation and Execution steps. This is because collection of a relevant corpus and re-training of the algorithm is needed locally before the phenotype algorithm may perform optimally. For the portability tradeoff model, we believe our proposed criteria relaxation and site-specific customization approaches would apply to ML-based phenotypes, as features used within the model may be removed (relaxed) from the model for all implementers, or require other local changes depending on factors such as data availability. More recent advances in ML have included the concept of “federated learning,” whereby predictive models may be trained across disparate data sources that are not centrally available to the model developer (e.g., to preserve privacy). This has been demonstrated in the healthcare space, and future work in this area will be an important consideration for portable ML-based phenotype algorithms that can automate the evolution of phenotype algorithm by iterating through all sites’ data.

We do acknowledge some limitations of this work. Although we are proposing our framework as a global one, we recognize that this is not based on a systematic literature review, and is focused on the health and biomedical domain. As our reconciliation of the network-specific models was based on multiple permutations, and we reached a saturation of new permutations during the review, we do believe the model represents the current state of the field. Similarly, as new knowledge is gained within the health and biomedical domain and beyond, the model can (and should) be adjusted accordingly. We also recognize that no evaluation of the reviewed research networks is presented using this framework. In development of the framework, we observed that heterogeneity within each network can and does exist (as previously mentioned). This precluded us from classifying each network as a whole, and instead requires formal study within and across the respective networks. We believe this is important future work to be undertaken, but was outside of the scope of this paper. Given current knowledge, this model provides health and biomedical informatics researchers with a framework by which to formally evaluate and further study aspects of portability of EHR-based phenotype algorithms.

Conclusion
Based on a review of seven multi-site research networks, we have identified multiple steps and considerations that impact how an EHR-based phenotype algorithm may be made portable. We have provided a more comprehensive definition of phenotype algorithm portability by proposing a framework with eight steps across three domains (Data, Authoring, Implementation). Within these domains, we have also identified and illustrated how optimizing one can result in more or less effort required in another, and how multiple representations for portable phenotype algorithms have been used. We believe this framework will help guide future research in the area of phenotype algorithm portability.

Acknowledgements
This work was supported in part by NIGMS grant R01GM105688, NHGRI grants U01HG008673 and U01HG6379, and NCATS grant UL1TR001422. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References
28. OHDSI. OHDSI/PhenotypeLibrary: A repository to store validated phenotype cohort definitions, with all associated source code, documentation, and validation results. 2019 [March 13, 2019]. Available from: https://github.com/OHDSI/PhenotypeLibrary.


Migrating from One Comprehensive Commercial EHR to Another: Perceptions of Front-line Clinicians and Staff

Tera L. Reynolds, MPH, MA1, Brian Clay, MD2, Scott E. Rudkin, MD, MBA3, Sara Beckham1, Danielle Perret Karimi, MD3, Joshua Glandorf2, Pat Patton, MSN, RN3, Christopher Longhurst, MD, MS3, Kai Zheng, PhD1

1University of California, Irvine, Irvine, CA; 2University of California San Diego Health System, San Diego, CA; 3University of California Irvine Health, Irvine, CA; 4Optimum Healthcare IT, Jacksonville Beach, FL

Abstract

As healthcare organizations continue to grow and evolve, migrations from one commercial electronic health record (EHR) system to another are likely to become more common. However, little is known about front-line clinicians’ and staff’s perceptions of such changes. Our study addresses this gap through an organization-wide survey of employees immediately prior to the transition to a new commercial EHR. We found that almost all front-line clinicians and staff were aware of the upcoming migration, and that most felt positive or neutral about the change, with only about 11% indicating that they were uncomfortable with the migration. Reasons for discomfort included the beliefs that the new EHR will be more time consuming to use and that moving to a new EHR is too costly, as well as concerns about the migration process. Attitudes differed by demographic characteristics and satisfaction with the current EHR. We discuss the implications of these results.

Introduction

While the investments made through the 2009 HITECH Act have resulted in significant progress in the adoption of electronic health record (EHR) systems in the U.S. over the last decade, this progress has come with some growing pains.1,2 Many healthcare organizations have now not only implemented an EHR but have also migrated from one EHR to another. The extant literature suggests that these migrations have met numerous challenges and that these challenges can put patient safety at risk.3 For instance, studies have found that technical issues (e.g., converting data to a new format) can create new work for users and result in continued parallel reliance on the old system.4–6 In addition, physicians often perceive such changes negatively, for example, reporting that the new EHR is less efficient.5–7 In fact, Hanauer et al. reported that, even up to two years after migrating to a popular commercial EHR system, many physicians’ attitudes towards the new EHR were surprisingly negative.5

Existing studies have largely focused on migrations from locally developed, “homegrown,” EHRs to commercial systems.6–7 In these cases, healthcare organizations are switching from an EHR that has been designed specifically for their particular context, but that may not be a comprehensive system with the more advanced functionalities offered by a mature commercial tool (e.g., clinical decision support). However, in recent years, it has become increasingly common for healthcare organizations to migrate from one commercial EHR to another commercial system. Like the aforementioned transitions, these migrations may be motivated by the need for a commercial tool with more advanced functionalities. In other cases, though, healthcare organizations have come to realize that the comprehensive commercial tool that they adopted is not optimal for their situation, prompting them to switch to a more suitable comprehensive commercial EHR.

Less is known about commercial-to-commercial migrations, especially how front-line clinicians and staff perceive the change.8,9 Our study addresses this gap through a survey assessing the attitudes and beliefs of employees at a large academic health system immediately prior to such a transition. The results of this survey have important implications for the migration at the study institution, as well as those considering or planning for a similar transition.

Methods

Institutional Setting

University of California, Irvine Health (UCI Health) is a large academic health system. In November 2017, UCI Health migrated from their existing comprehensive commercial EHR, Quest (Allscripts, Chicago, IL, USA), to another comprehensive commercial EHR, Epic (Epic Systems, Verona, WI, USA). UCI Health prepared clinicians and staff for this change in a number of ways such as in-person and online trainings.
**Survey Development and Deployment**

We developed a short survey to understand clinician and staff perceptions related to the upcoming migration from Quest to Epic. Specifically, we included structured questions to understand their satisfaction with Quest, awareness of and comfort with the upcoming migration to Epic, and reasons for their attitudes. Respondents were also able to provide unstructured (free-text) responses to several of the questions. The technology acceptance model (TAM) was used to guide survey development.

To minimize the burden on participants, we designed a skip pattern that ensured that they were only asked to respond to questions that were appropriate based on their previous answers. For instance, if a participant indicated that they did not know what Quest was, they were not asked to complete any additional questions. After the survey instrument was finalized, the questions and survey logic were entered into REDCap. The initial invitation email was sent through REDCap to 7,361 UCI Health employees on September 12, 2017. Two reminder emails were subsequently sent to non-respondents over the five weeks the survey remained open.

**EHR Account Status and Demographics**

In addition to the survey results, we also had an EHR account status dataset and a demographics dataset for all UCI Health employees. The former dataset included an identifier, Epic account status (active or inactive), date of last access, and provider type. The latter included the following variables: an identifier, department, years at UCI Health (as of the date they took the survey or, for non-respondents, September 18, 2017), FTE, shift, job title, and job family (i.e., job type).

UCI Health categorizes career opportunities into the following job families: (i) administrative/clerical (e.g., clerks, administrative assistants), (ii) clinical professional (e.g., case managers, social workers), (iii) executive (e.g., directors), (iv) finance/information technology (e.g., programmers), (v) management (e.g., administrative nurses), (vi) nursing (e.g., clinical nurses), (vii) professional (e.g., administrative analysts), (viii) support services (e.g., medical assistants, admitting workers), (ix) technician/technologist (e.g., anesthesia technicians), and (x) trades and services (e.g., mechanics). We separated out clinical support service (e.g., medical assistants) and other support service (e.g., medical records admin) jobs to better align with the other job families and because we believed that those with a patient care role may have a different perspective than those with a more administrative role. In addition, 2,881 records were missing the job family field, including all physicians, as well as those with a professor or instructor job title. Therefore, we created two new job families: physician and non-clinical professor/instructor. The latter category only included professors or instructors without an active EHR account, indicating that they are not practicing clinicians. After doing this, only 472 were missing the job family field. These were categorized into the existing job families as appropriate. For example, 98.8% of job titles that included the word ‘supervisor’ were associated with the management job family; thus, if the job title for a record without a job family included this term, it was categorized as management.

In addition to the existing variables, we also created a new variable, which we refer to as job level. We categorized employees into one of five different job levels based on their job title and/or job family: (i) trainee (e.g., ‘resident’ or ‘intern’ in job title), (ii) early-career/entry (e.g., ‘i’ or ‘ii’ in job title as in ‘clinical nurse ii’), (iii) mid-career (e.g., ‘iii’, ‘iv’, or ‘v’ in job title as in ‘biller iv’), (iv) senior (e.g., ‘lead’ in job title), and (v) leadership (e.g., ‘executive’ job family).

**Data Analysis**

We combined the three datasets for analysis. Structured data were analyzed using R v.3.5.1. After conducting univariate analyses, we assessed the representativeness of our sample. We then conducted bivariate analyses to evaluate whether survey responses varied by Epic account status and/or the demographic variables, as well as to determine whether responses to certain survey questions were associated with other survey responses. For instance, we hypothesized that satisfaction with Quest would be inversely related to comfort with the migration to Epic. In other words, those dissatisfied with Quest would be more likely to be comfortable with the migration to Epic than those satisfied with or neutral about Quest (and vice versa).

For categorical independent variables we used chi-square tests for independence of all factors. We then used adjusted standardized residuals to identify where the significant differences occurred. To correct for multiple comparisons, we applied the Bonferroni p-value correction method. In addition, for continuous independent variables we performed pairwise comparisons using t-tests with pooled standard deviations, and the Holm-Bonferroni p-value correction method to account for multiple comparisons.
Free-text survey responses were analyzed with NVivo v.12.2.0. We employed an inductive qualitative analysis approach and the constant comparison method to identify emergent categories and then to classify participants’ comments.10,11 We continued this process until no new categories emerged (theoretical saturation).12 This was a quality improvement study and was, thus, exempt from institutional review board review.

**Results**

Of those invited, 1,674 responded to the survey (22.7%). Nine did not fill out the first question; and 47 indicated that they did not know what is “Quest.” These 56 individuals were excluded from the analysis. Unsurprisingly, among those excluded, the majority were not active EHR users (60.7%, N=56). In addition, a relatively high percentage of respondents in non-clinical jobs were excluded from analysis, including non-practicing clinical professors/instructors (Excluded: 20%, N=20) and trades and services jobs (Excluded: 14.3%, N=14). On the other hand, responding physicians (0%, N=222), finance professionals (0%, N=45), executives (0%, N=26), and nurses (0.7%, N=401) were least likely to be excluded.

Relative to the entire population of UCI Health employees, physicians are underrepresented in the analyzed sample (Analyzed: 13.7% vs. Overall: 20.6%), while nurses (Analyzed: 24.6% vs. Overall: 18.9%) and managers (Analyzed: 10.4% vs. Overall: 5.0%) are overrepresented. Furthermore, while the percentage of respondents in early-career or entry-level, mid-career, and senior positions are similar to the overall population, a relatively lower percentage of trainees (e.g., residents) were included in the analysis (Analyzed: 5.1% vs. Overall: 10.8%).

**Satisfaction with the Current EHR**

Among the analyzed sample (N=1,618), most were satisfied with Quest (45.9%), the EHR system currently used at UCI Health, or were neutral (31.0%). However, there were differences in satisfaction based on the respondent’s job type ($\chi^2=133.41$, df=16, p<0.001). Figure 1 presents a summary of these results. Upon examining the results of the post-hoc analysis ($\alpha_{adj}=0.002$, $z$ criteria=-3.11), we found that significantly more physicians were dissatisfied with Quest than would be expected if there was no association between the two variables, and significantly fewer were satisfied than expected (N=222; Dissatisfied: 50.5%, $r_{adj}=10.46$; Satisfied: 24.8%, $r_{adj}=-6.81$).

![Figure 1. Satisfaction with the current EHR, Quest, by job type.](image-url)
There was also a statistically significant association between satisfaction with Quest and job level ($\chi^2=41.61$, df=8, p<0.001). Specifically, the results show that significantly more trainees (e.g., residents) were dissatisfied with Quest than expected and significantly fewer were satisfied than expected ($\alpha_{adj}=0.003$, z criteria=-2.94; N=82; Dissatisfied: 48.8%, $r_{adj}=5.69$; Satisfied: 22.0%, $r_{adj}=-4.48$). Along these same lines, we also found that, on average, those satisfied with Quest had been employed at UCI Health longer than those who were dissatisfied with Quest (Satisfied: Mean=10.2 years, SD=7.98 vs. Dissatisfied: Mean=8.78 years, SD=7.74; p= 0.016).

### Awareness of the Migration

The vast majority of respondents, 98.5%, were aware that UCI Health would soon replace the current EHR, Quest, with Epic, another electronic health record system (N=1,612). The high level of awareness was fairly uniform across demographic variables, with only negligible differences.

### Attitude Toward the Migration

When asked about how they feel about UCI Health’s plan to replace Quest with Epic, most respondents indicated that they were comfortable with the migration (54.4%, N=1,601), with only 10.9% saying that they were uncomfortable. The remainder indicated that they felt neutral about the migration (34.7%).

There was some variation in responses based on demographic variables, including type of job ($\chi^2=78.6$, df=16, p<0.001) and job level ($\chi^2=66.15$, df=8, p<0.001). The post-hoc analysis showed that, for job type ($\alpha_{adj}=0.002$, z criteria=3.11), significantly more management/executives were comfortable with the migration than expected and significantly fewer than expected were neutral (N=190; Comfortable: 69.5%, $r_{adj}=4.44$; Neutral: 17.9%, $r_{adj}=-5.19$). In addition, significantly fewer physicians were neutral about the migration than expected (21.8%, N=220, $r_{adj}=-4.33$), while significantly more technicians/technologists were neutral about the migration than expected (53.8%, N=104, $r_{adj}=4.23$).

For job level ($\alpha_{adj}=0.003$, z criteria=-2.94; See Figure 2), we found that the results for the leadership category were very similar to the management/executive job type category. We also found that significantly fewer in mid-career were comfortable with the migration than expected (44.2%, N=231, $r_{adj}=-3.38$). On the other hand, significantly more trainees were comfortable with the migration than expected and significantly fewer than expected were neutral (N=81; Comfortable: 74.1%, $r_{adj}=3.65$; Neutral: 11.1%, $r_{adj}=-4.58$).

![Figure 2. Attitude toward the Epic migration by job level.](image-url)
There was also a statistically significant difference in attitude toward the Epic migration based on satisfaction with Quest ($\chi^2=110.79$, df=4, $p<0.001$). These results are highlighted in Figure 3. Perhaps unsurprisingly, the post-hoc analysis ($\alpha_{adj}=0.006$, $z$ criteria=-2.77) revealed that there were significantly more participants who were dissatisfied with Quest and comfortable with the migration than expected (73.4%, $N=368$, $r_{adj}=8.32$), while there were significantly fewer individuals satisfied with Quest and comfortable with the migration than expected (43.5%, $N=738$, $r_{adj}=-8.10$). In addition, participants who reported being satisfied with Quest and uncomfortable with the migration were observed significantly more than expected (15.3%, $N=738$, $r_{adj}=5.28$) and those neutral about Quest and uncomfortable with the migration were observed significantly less than expected (5.5%, $N=495$, $r_{adj}=-4.66$).

Figure 3. Attitude toward the Epic migration by satisfaction with the current EHR, Quest.

Those who felt comfortable with migrating to Epic ($N=871$) were asked to provide a reason(s) for this perception (see Table 1). Among the 858 participants that selected at least one of the options, over half indicated that they believe that Epic will better support their clinical work (60.4%), will be less time consuming to use (58.7%), will make it easier to retrieve data (57.6%), and that Epic’s user interface is easier to learn and use (54.1%).

Those who felt uncomfortable with migrating to Epic ($N=174$) were also asked to provide a reason(s) why they felt this way (see Table 1). Although none of the response options were selected by fifty percent or more of the 153 participants that chose at least one reason, the four most commonly reported attitudes were that Epic will be more time consuming to use (41.8%), that moving to a new electronic medical records system is too expensive (38.6%), that Epic’s user interface is more difficult to learn and use (29.4%), and that Epic will not support their clinical work as well (24.8%).

Table 1. Respondents’ reasons for feeling comfortable or uncomfortable with the migration to Epic.

<table>
<thead>
<tr>
<th>Reason for Feeling Comfortable</th>
<th>Percentage of Participants ($N=858$)</th>
<th>Reason for Feeling Uncomfortable</th>
<th>Percentage of Participants ($N=153$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epic will better support my clinical work</td>
<td>60.4%</td>
<td>Epic will not support my clinical work as well</td>
<td>24.8%</td>
</tr>
<tr>
<td>Epic will be less time consuming to use</td>
<td>58.7%</td>
<td>Epic will be more time consuming to use</td>
<td>41.8%</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Moving to a new electronic medical records system is too expensive</td>
<td>38.6%</td>
</tr>
</tbody>
</table>
In addition to the structured questions, participants had several opportunities to share unstructured comments. These free-text responses revealed additional insights into participants’ attitudes towards the migration to Epic and fell into two main categories – attitudes towards (1) the technologies (Quest and/or Epic) and (2) the migration process. First, as suggested by the quantitative results, many participants were positive about transitioning to Epic. This was generally either due to negative perceptions about Quest or positive perceptions about Epic. For instance, one person with little knowledge about Epic stated, “It couldn't be any worse [than Quest].” Other participant comments focused on their positive beliefs or expectations about Epic, including that it will improve coordination and communication; has better system design, functionality, or usability; offers a process efficiency or otherwise better supports their work; or will improve outcomes such as increased patient safety (see Table 2). Participants indicated that their positive beliefs about Epic were largely based on their knowledge of the EHR market (e.g., “it is an industry standard”), personal experience with Epic at another institution, co-workers’ experiences with Epic, or organizational trust (e.g., “I am sure UCI has done enough research, and if it shows medical record improvement with Epic, why not”).

Second, while some participants were not actually opposed to migrating to Epic, they did feel that certain aspects of the migration process had been problematic, including training, planning, communication, and leadership (see Table 2). Of these, insufficient training and the perception of poor planning were the most commonly mentioned. Several respondents felt that they were not adequately trained to use Epic and, thus, felt ill prepared for the upcoming change. For instance, one participant stated, “training end users with incomplete workflows that can impact patient care patient's financial status. I do not feel we are ready for Go Live…” In addition, issues with the migration plan, especially the timeline, were also relatively common. For example, one person stated, “Concept is good, but timeline is too aggressive…” Finally, some respondents cited insufficient communication and negative perceptions of the migration leadership as sources of uncertainty or concern. For example, one participant stated, “EPIC’s functionality and how it differs from Quest has not been made known to us…” Another respondent commented, “The consultants that were hired to direct this project are ineffective.”

Table 2. Summary of free-text survey results.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technologies</strong></td>
<td></td>
</tr>
<tr>
<td>Negative Perceptions about Current EHR</td>
<td>“I don't know enough about Epic to know whether any of the above [reason feel comfortable question options] are true, but there's no way that it won't be better than Quest.”</td>
</tr>
<tr>
<td></td>
<td>“Currently, we use 3 to 4 different systems for our patients.”</td>
</tr>
</tbody>
</table>
• “I believe Epic it's from this century, not like Quest.”

### Positive Perceptions about New EHR

**Improves Coordination and Communication**
- “Will have access of patient medical records chart across EPIC system”
- “Epic will make coding easier in regards to work assignment…”

**Better System Design, Functionality, or Usability**
- “better reporting - real time”
- “Epic gives more control to the user by more versatile”
- “better and faster note template and loading”

**Offers Process Efficiency or Otherwise Better Supports Work**
- “Automatically generates cost vs me having to manually input billing”
- “Epic does not use the same Script as QUEST and the interface is set up like a Word document as opposed to the windows seen in QUEST so multitasking by writing notes and looking up labs is easier and more efficient”

**Improve Outcomes**
- “After Epic implementation, Epic alone will include these 3 to 4 systems to where nursing and radiology are using the same system, decreasing duplication of work, and increasing patient safety.”

### Migration Outcomes

**Insufficient Communication**
“EPIC’s functionality and how it differs from Quest has not been made known to us…”

**Insufficient Training**
“the training had an incomplete buildout and still a lot of unanswered questions. We need more training than is supplied in the on-line modules and the single face to face session.”

**Negative Perceptions about Migration Leadership**
“The consultants that were hired to direct this project are ineffective.”

**Problems with Planning**
- “I think the push to meet a target date is a little foolhardy. We should make sure the system works first, not worry about the calendar.”
- “Why does administration think that the conversion from Quest to Epic will not impact our clinical practice and that we do not need to adjust patient volumes?”

### Discussion

Through a large survey of front-line clinicians and staff at UCI Health, we found that almost all were aware of the upcoming migration to a new comprehensive commercial EHR, and that most were fairly positive about this change. However, there was some variation in attitudes towards the migration based on satisfaction with the current comprehensive commercial EHR. In addition, we found that satisfaction with the current EHR varied based on job type, with a higher proportion of physicians being dissatisfied compared to other types of jobs. Qualitative results provided additional insights showing that positive attitudes towards the migration tended to either be based on dissatisfaction with the current EHR or on positive beliefs or expectations about the new system, while negative attitudes were not always related to the technology but, rather, sometimes stemmed from dissatisfaction with the migration process.

Overall, our results are more positive than the findings from studies that have focused on migrations from homegrown EHRs to commercial systems, e.g.,[4-7] This may suggest relatively lower levels of clinician and staff resistance to transitions from one comprehensive commercial EHR to another. There could be a number of explanations for this. For instance, in homegrown-to-commercial migrations, clinicians’ loyalty to the homegrown system that was created specifically for their context e.g.,[6] and/or a general preference for simpler EHRs without advanced features e.g.,[13,14] could lead to resistance to the migration. While in the case of comprehensive commercial-to-comprehensive commercial
transitions, on the other hand, clinicians and staff may not have the same loyalty to the existing system and have already adjusted to a comprehensive EHR, which could result in lower levels of resistance to change.

With that said, however, our results suggest that more could have been done to facilitate a smooth transition to the new EHR, including (1) supporting key clinical activities; (2) addressing commonly cited reasons for discomfort, and (3) improving training. These findings have implications for other healthcare organizations that are considering or planning for a similar change.

First, both among those who felt comfortable with migrating to a new EHR and those who felt uncomfortable with this change, most of the commonly reported reasons were related to day-to-day clinical practice as opposed to larger institutional goals such as supporting quality improvement. Others have suggested that misaligned clinical and organizational needs can contribute to clinician dissatisfaction with EHRs. However, thus far, such migrations to be successful, the new EHR needs to adequately support key clinical activities, and clinicians and staff need to be reassured of this fact (e.g., through trainings and other communications).

Second, healthcare organizations should also specifically address the commonly reported reasons that clinicians and staff feel uncomfortable with a migration. For instance, clearly communicating the differences between the current EHR and new EHR and why the investment in the new EHR is necessary, as well as incorporating clinician and staff suggestions to improve the migration process, could help to mitigate many of the concerns. In addition, it may be necessary to tailor messages about the migration to different groups who may be more resistant to the change (e.g., those who are more likely to be satisfied with the current EHR).

Finally, and as many others have noted, adequate end-user training and on-going support is essential when implementing a new EHR. However, many of our participants perceived the training they had received thus far as insufficient—their participants wanted more personalized sessions and more opportunities to interact with the new EHR. Although there are some published training programs that end-users have found satisfactory, more research is needed to understand which components are most helpful and how trainings can be effectively adapted for different contexts.

Although this survey was conducted at a single institution, it included a number of different perspectives, including physicians (from both inpatient and ambulatory care), nurses, other clinical professionals, and staff. Further research is needed to validate our findings across other institutions. In addition, future studies should consider a longitudinal approach to understand how front-line clinician and staff perceptions change once they have used the new comprehensive commercial EHR in practice. Existing research suggests that, in order to see physician satisfaction rise above baseline, follow-up may need to occur more than two years after migrating from a homegrown EHR to a commercial system. However, it remains to be seen if that will be the case for comprehensive commercial-to-comprehensive commercial migrations.

Conclusion

We found that, immediately prior to the transition, front-line clinicians and staff were more positive about the migration from one comprehensive commercial EHR to another than had been reported in the literature on migrations from homegrown EHRs to commercial systems. However, there were still some employees who were uncomfortable with the migration. Addressing the reasons for this discomfort could reduce resistance and increase satisfaction with the new EHR once it is implemented. Further research is needed to understand how clinician and staff perspectives change after this type of migration.

References


Public Attitudes Toward Direct to Consumer Genetic Testing

Grayson L. Ruhl1, James W. Hazel, PhD, JD1, Ellen Wright Clayton, MD, JD1, Bradley A. Malin, PhD1
Vanderbilt University, Nashville, Tennessee, USA

Abstract
Direct to consumer genetic testing (DTC-GT) is an emerging service that allows individuals to have their DNA tested without having to consult a healthcare provider. DTC-GT can provide insight into various aspects about an individual, including their health and ancestry. However, testing may pose privacy risks and yield distressing results. Despite the growing popularity of DTC-GT, public attitudes toward such services remain largely ill-defined. Using Amazon Mechanical Turk, we administered a web-based survey to over 1,000 individuals to obtain intuition into public attitudes about DTC-GT. Survey questions were grounded in a literature review of people’s views about DTC-GT. The results of the survey indicated that respondents were interested in DTC-GT as a possible way to gain insight about health, ancestry, and family relationships, as well as advance research. Despite this, respondents were concerned that DTC-GT companies and other users of their DTC-GT data would infringe upon their privacy.

Introduction
Advances in genome sequencing technology have enabled rapid growth in genomics research since the first sequencing of the human genome. As the price of genome sequencing has dropped over time, and the technology has become more accessible to the public, direct-to-consumer genetic testing (DTC-GT) has emerged as a new service. DTC-GT, also known as at-home testing, entails genetic testing services that are marketed directly to consumers and enable individuals to obtain personal genetic information without having to consult a healthcare provider. For example, 23andMe.com sends test kits to consumers’ homes to collect DNA samples in the form of saliva. Consumers then send the kits to a laboratory and receive results online after about six to eight weeks.

Research studies in this area, which are largely focused on well-established and popular DTC-GT companies such as 23andMe.com and Ancestry.com, have indicated that there are several factors that directly influence consumer decisions to undergo testing. While DTC-GT can offer insight about a wide variety of topics, many people are particularly drawn to DTC-GT services that claim to provide information about personal health (e.g., informing participants about genetic health risks and carrier status), wellness, ancestry, and personal traits. Some consumers have been found to trust DTC-GT companies more than academic medical centers for the purpose of obtaining personal genetic information. Some services, however, are more controversial than others. Some companies offer tests that purport to assess children’s athletic potential, covertly determine paternity, and discover infidelity by testing surreptitiously-gathered biological material. Such practices have raised a variety of concerns among the ethical, legal, and scientific communities. This new market for genomics services has stirred debates about genetic privacy and identity, as there are potential risks and ethical issues associated with DTC-GT.

The current literature about DTC-GT indicates that potential consumers are, to some degree, concerned about a variety of risks. Concerns include, but are not limited to, the following: (a) lack of certainty regarding how much control they have over data, (b) worries about privacy invasion, (c) tests that reveal health complications, determine paternity, or expose infidelity may disrupt family relationships or otherwise cause distress, and (d) misunderstandings due to inability to understand results. Despite these worries, this space is largely unregulated, except by the Food and Drug Administration (FDA) for health-related results. Recent studies of companies offering DTC-GT testing demonstrate that their privacy policies often fail to meet transparency guidelines concerning privacy and secondary use.

In light of these concerns, consumer enthusiasm for DTC-GT can be difficult to rationalize. The privacy paradox is a commonly observed phenomenon wherein one’s behavior concerning privacy risks does not reflect his or her privacy concerns. Consumers often succumb to this paradox when weighing the utilities of having personal genetic testing against the entailed risks, and they are often unaware of potential consequences. Furthermore, preferences and actions vary by the individual, complicating our understandings of what drives people to pursue or avoid testing. Due to a lack of public understanding about this new arena of genetic testing, this market lacks policies that would meet consumers’ needs while ensuring the delivery of useful results and mitigating the problems that testing could cause.
Insight into how individuals react to the influences surrounding personal genetic testing will be useful in forming policies to create an optimal environment for participation in DTC-GT. This would be mutually beneficial to consumers and genetic testing companies. Specifically, consumers would experience the benefits of personalized genetic testing while facing low associated risks. Additionally, DTC-GT would become a more popular and transparent service industry. Informed by a review of the literature surrounding DTC-GT and a comprehensive analysis of DTC-GT company advertisements, we created a survey to understand public attitudes toward DTC-GT in greater depth.

**Methods**

**Goals and Literature Review**

The overarching goal of this study was to decipher what individuals perceived as the utilities and risks of having DTC-GT performed. To realize this goal, we 1) codified a grounded theory about public attitudes toward DTC-GT, which was inspired by the literature surrounding the DTC-GT landscape; 2) designed and administered a survey based on these observations; and 3) tested the survey results against the grounded theory. The study was approved by Vanderbilt University’s Institutional Review Board (Protocol #180172) and data collection took place in November 2018.

To gain an understanding of the DTC-GT environment, we searched for studies that explored the DTC-GT landscape. Having collected and examined a comprehensive set of studies that specifically assessed public attitudes toward DTC-GT services, we developed theories about the motivators that contribute to peoples’ opinions, as described below. We also researched DTC-GT advertisements by visiting the websites of various companies known to the investigators and by viewing commercials aired by companies. After conducting the literature review, our grounded theory was formulated as follows:

a) People view DTC-GT as a useful means for managing their health and learning about their ancestry, but also utilize it for altruistic purposes (e.g., helping a family member or advancing research).6,19

b) Peoples’ most prevalent concerns surrounding DTC-GT are 1) privacy risks;12 2) the interpretability and emotional impact of results;19,20 and 3) the familial disruptions that various types of results would cause.13

c) People are more comfortable sharing their DTC-GT results with academic institutions than with third parties, including governmental, commercial, or foreign entities.7

Gauging privacy concerns required deliberation: privacy is a vague concept that is difficult to articulate and can embody various meanings. To determine public attitudes about the privacy risks while avoiding confusion, we separated privacy into three applicable concepts: confidentiality, anonymity, and solitude.21 Confidentiality entails keeping one’s information restricted to certain entities. Anonymity entails protecting one’s identity. And solitude entails respecting one’s right to be left alone, shielded from unwanted solicitation. We designed questions that would measure respondents’ concerns that DTC-GT companies would breach these three aspects of privacy. We were also interested in how social relationships influenced in consumers’ decisions. We assessed this by asking respondents about how motivated they would be to undergo testing if certain social connections (e.g., their parents, children, and friends) thought it was a good idea.

**Survey Creation**

Our survey instrument was created using the Research Electronic Data Capture (REDCap) platform22 and was administered on the Amazon Mechanical Turk (MTurk; discussed below). The REDCap survey consisted of various types of multiple-choice questions regarding DTC-GT, 15 in total. Respondents were also asked to provide voluntary demographics information (see Table 1). To measure respondents’ attitudes, we used the 3-point Likert scale answers, “A lot,” “Somewhat,” and “Very little” (the phrasing of some of our questions required slight wording variations for these Likert scale options). Several questions within our survey addressed topics using multiple choice inquiries, but not through a Likert scale, which allowed us to investigate connections between the Likert scale scores and the factors that potentially contributed to respondents’ attitudes. To verify participation and facilitate compensation, the survey also contained a field wherein respondents entered their MTurk worker IDs.

To determine the validity of workers’ responses before sending out our final survey, we gathered responses from a sample of 50 individuals. The responses from this sample demonstrated that certain questions contained wordings that resulted in misinterpretation, which prompted efforts to clarify the meanings of each question. To achieve face validity, we solicited feedback on the design and content of our survey from colleagues within Vanderbilt’s Center for Genetic Privacy and Identity in Community Settings (GetPreCiSe), a diverse interdisciplinary working group consisting of scholars from the law, medicine, psychology, history, and humanities fields. As further validation, and to ensure clarity
of our questions after these extensive refinements, we conducted cognitive interviews with a small group of undergraduate students at Vanderbilt University. During these interviews, participants were asked to read aloud each question and provide their assessment of its intent. In response to this feedback, we concluded that the survey questions were unambiguous and comprehensible.

**Study Population**

The population for this investigation consists of respondents who took our survey, according to our guidelines, through the Amazon Mechanical Turk (MTurk) online crowdsourcing platform. MTurk is Amazon’s online marketplace for individuals, often called workers, to complete various tasks (e.g., transcribing a speech, determining if a video contains explicit material, or answering surveys), usually for some money in return. MTurk served as a portal to prompt respondents about our survey and to compensate them $0.50 for their participation. On MTurk’s interface, we provided background information about DTC-GT, as well as a consent form, and provided respondents with the link and code necessary to access our REDCap survey.

**Survey Dissemination and Data Analysis**

We then released our survey, making it available to workers who resided in the United States for four days. After requesting 1,050 responses through MTurk, we gathered 1,090 surveys on REDCap. Ultimately, 1,026 submissions were deemed viable. We removed 64 from consideration for the following reasons: 34 submissions were unique to the REDCap system and thus did not coincide with MTurk Worker IDs; 24 submissions contained duplicates; and 6 submissions contained responses with no variance in their Likert scale answers, which strongly suggested that these workers did not consciously participate in our survey.

We exported the raw data files from REDCap containing the responses of all respondents mapped to a numerical format and derived general summary statistics about the results. For instance, answers for Likert scale questions showed up as a “1” to represent “A lot,” a “2” to represent “Somewhat,” and a “3” to represent “Very little.”

**Results**

**Respondents’ Participation in DTC-GT**

Of all respondents, 178 (17.3%) had undergone DTC-GT, 821 (80.0%) had not, and 27 (2.6%) were unsure (See Figure 1). Of the 178 respondents who had testing done, 117 (65.7%) did so to learn about their ancestry, 66 (37.1%) were curious about their personal traits, 58 (32.6%) wanted help managing their health, 19 (10.7%) wanted to determine the paternity of a child, and 5 (2.8%) took a DTC-GT test to discover infidelity (See Figure 2). Of the 821 respondents who had not undergone testing, 346 (42.1%) said they were likely to have testing in the future, 247 (30.0%) said they were unlikely to, and 228 (27.8%) did not know whether they would or not.

**Perceived Utilities of DTC-GT**

Of all respondents, 90 (8.8%) knew more than five friends or family members who had undergone DTC-GT, 471 (45.9%) knew at least one such person, 256 (25.0%) knew none, and 209 (20.4%) did not know. When asked about the extent to which certain social connections would motivate them to get testing (i.e., if these individuals or groups

---

a Here we described the purpose, approximate duration (five to ten minutes), expected costs, risks and benefits, content, and expected compensation of our study.

b We removed these 34 submissions after having deleted duplicates and incomplete submissions. Furthermore, a few of the MTurk entries had no REDCap submissions to match them. These respondents were not paid.

c We encountered repeated cases of Worker IDs within the exported REDCap dataset, suggesting that workers took the survey twice. In these cases, we removed all surveys that an individual submitted after their first submission.

d Every question required an answer so that we could conduct thorough and consistent analysis; surveys that were either exited or not completed were not considered, and we did not pay respondents who submitted such results.

e We were able to determine the legitimacy of MTurk worker responses by inspecting the relationship between average Likert scale scores and their standard deviations. We assumed relatively high standard deviations indicated that respondents were answering questions thoughtfully.
thought it was a good idea), respondents’ spouses or partners were most influential (362 [35.3%] said “A lot”), followed by their children (282 [27.5%] said “A lot”), parents (197 [19.2%] said “A lot”), and friends (170 [16.6%] said “A lot”) (See Figure 3). There were 401 (39.1%) respondents who said that their friends’ positive opinions of DTC-GT would motivate them very little.

Of all respondents, 707 (68.9%) said that they would be strongly motivated to have DTC-GT done if it would help them in combating their own illness, and 664 (64.7%) would be strongly motivated if it would help someone in their family with a healthcare issue. When asked how motivated they would be to pursue testing if it would help advance medical research, 430 (41.9%) respondents said “A lot,” and 372 (36.3%) said “Somewhat.” There were 538 (52.4%) respondents who answered that reports in the news or social media that made DTC-GT seem like a good idea would influence them very little (See Figure 4).

When asked how much they thought DTC-GT could provide useful or interesting information about certain topics, 568 (55.4%) said “A lot” about their ancestry, 471 (45.9%) said “A lot” about family relationships (e.g., paternity), 404 (39.4%) said “A lot” about managing their health, 281 (27.4%) said “A lot” about their personal traits, and 172 (16.8%) said “A lot” about infidelity (See Figure 5). More than half (526 [51.3%]) thought that DTC-GT could provide very little information about infidelity.

**Concerns about DTC-GT**

When asked how much control individuals should have over who may use their DTC-GT data, 809 (78.8%) respondents answered “A lot,” 177 (17.3%) said “Some,” and 40 (3.9%) answered “Very little.” Figure 6 illustrates respondents’ concerns about anonymity breaches, confidentiality breaches, and infringement upon solitude. Of all participants, 507 (49.4%) were very concerned that DTC-GT companies would fail to protect their identity, 489 (47.7%) said the same about sharing information without their consent, 428 (41.7%) said the same about providing unsolicited advertisements or emails, and 391 (38.1%) said the same about providing unsolicited health information.

**Figure 2.** Responses to the question, “Why did you decide to take a DTC genetic test?”

<table>
<thead>
<tr>
<th>Motivation</th>
<th># of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>It would help in combating your own illness</td>
<td>777</td>
</tr>
<tr>
<td>It would help someone in your family with a healthcare issue</td>
<td>777</td>
</tr>
<tr>
<td>You think it will help advance medical research</td>
<td>777</td>
</tr>
<tr>
<td>Reports in the news or social media made it seem like a good idea</td>
<td>777</td>
</tr>
</tbody>
</table>

**Figure 3.** Responses to the question, “How motivated would you be to have DTC genetic testing if your ___ thought it was a good idea?”

**Figure 4.** Responses to the question, “How motivated would you be to have DTC genetic testing if…”
Respondents also reported concerns about how downstream users (e.g., pharmaceutical companies or university research centers) might use their data. Of the respondents, 429 (41.8%) were very concerned, and 468 (45.6%) were somewhat concerned, that these users of their DTC-GT results would fail to protect their privacy (See Figure 7).

When asked whether users of their DTC-GT data should report new results (i.e., results generated beyond their initial testing) back to participants if such results indicated health concerns (e.g., an increased risk of developing breast cancer or an increased risk of addiction), 606 (59.1%) said “Always”, 378 (36.8%) indicated that they should report only serious health risks back, and 42 (4.1%) said “Never” (See Figure 8).

Respondents generally felt fairly confident that they could understand the results of DTC-GT. When asked to assess their ability to comprehend results, 275 (26.8%) said that they would accurately understand their results without any help from an expert; 558 (54.4%) said that they would understand some of their results but would prefer help from an expert; 134 (13.1%) said that they would need an expert or genetic counselor to help them understand their results; and 59 (5.8%) respondents were unsure about their capabilities to understand DTC-GT results.

Respondents were also asked how concerned they were that DTC-GT would reveal to them that they had a greater chance of developing an incurable disease, such as Alzheimer’s disease. Of all respondents, 180 (17.5%) were very concerned, 466 (45.4%) were moderately concerned, and 380 (37.0%) were not very concerned.

Given various types of DTC-GT results, respondents were then asked how concerned they were that they would disrupt their family relationships. Such concerns were generally low: 150 (14.6%) said “A lot” for health results, 105 (10.2%) said “A lot” for ancestry results, 98 (9.6%) said “A lot” for results about their personal traits, 100 (9.7%) said “A lot” for paternity results, and 93 (9.1%) said “A lot” for infidelity results (See Figure 9).
Use of DTC-GT Results

Respondents were asked with which groups they were comfortable sharing their DTC-GT results (e.g., to support research, for their personal care, or some other purpose). They were asked to mark all that applied: 646 (63.0%) respondents felt comfortable sharing results with academic institutions, 371 (36.2%) with third party companies required to provide the same privacy protections as the DTC-GT company, 147 (14.3%) with the government, 63 (6.1%) with commercial entities like Google or IBM, and 57 (5.6%) with people outside of the U.S. (see Figure 10).

When asked whether or not they would be willing to allow the results of DTC-GT to be used in research, 241 (23.5%) respondents said yes, even if their results were fully identified as theirs; 646 (63.0%) said yes, but only if their results were stripped of any identifying information; and 139 (13.5%) said “Never.”

---

**Figure 9.** Responses to the question, “How concerned are you that genetic test results about the following would disrupt your family relationships?”

**Figure 10.** Responses to the question, “With which of the following groups are you comfortable sharing your DTC genetic testing results?”

**Figure 11.** Responses to the question, “To what extent are you concerned that law enforcement can access peoples’ DTC-GT test results?”

**Figure 12.** Responses to the question, “To what extent should law enforcement have access to people’s DTC genetic test results to use for investigative purposes?”
Respondents were asked about their concerns surrounding law enforcement access to DTC genetic testing results. Of all respondents, 375 (36.5%) reported “A lot” of concern, 413 (40.3%) were “Somewhat” concerned, and 238 (23.2%) reported “Very little” concern (Figure 11). Respondents were also asked about the extent to which they felt law enforcement should have access to people’s DTC genetic test results to use for investigative purposes, either before or after a crime has been committed. Respondents appeared to be much more permissive of law enforcement accessing DTC genetic tests results to solve a crime than to prevent a crime: 734 (72.4%) of respondents said that law enforcement should have “very little” access to DTC genetic test results before a crime had been committed, while only 270 (26.3%) said that law enforcement should have “very little” access after a crime had been committed (Figure 12).

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Group</th>
<th># of Respondents</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>18 – 35</td>
<td>605</td>
<td>58.97%</td>
</tr>
<tr>
<td></td>
<td>36 – 59</td>
<td>345</td>
<td>33.63%</td>
</tr>
<tr>
<td></td>
<td>60 and over</td>
<td>51</td>
<td>4.97%</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>25</td>
<td>2.44%</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>549</td>
<td>53.51%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>463</td>
<td>45.13%</td>
</tr>
<tr>
<td></td>
<td>Non-binary</td>
<td>4</td>
<td>0.39%</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>10</td>
<td>0.97%</td>
</tr>
<tr>
<td>Hispanic, Latino, or</td>
<td>Yes</td>
<td>108</td>
<td>10.53%</td>
</tr>
<tr>
<td>Spanish origin</td>
<td>No</td>
<td>907</td>
<td>88.40%</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>11</td>
<td>1.07%</td>
</tr>
<tr>
<td>Race</td>
<td>African-American</td>
<td>90</td>
<td>8.77%</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>65</td>
<td>6.34%</td>
</tr>
<tr>
<td></td>
<td>Native American</td>
<td>19</td>
<td>1.85%</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>804</td>
<td>78.36%</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>35</td>
<td>3.41%</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>13</td>
<td>1.27%</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Single (never married)</td>
<td>476</td>
<td>46.39%</td>
</tr>
<tr>
<td></td>
<td>Married or in a domestic partnership</td>
<td>460</td>
<td>44.83%</td>
</tr>
<tr>
<td></td>
<td>Divorced</td>
<td>56</td>
<td>5.46%</td>
</tr>
<tr>
<td></td>
<td>Separated</td>
<td>10</td>
<td>0.97%</td>
</tr>
<tr>
<td></td>
<td>Widowed or widower</td>
<td>15</td>
<td>1.46%</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>9</td>
<td>0.88%</td>
</tr>
<tr>
<td>Religion</td>
<td>Very religious</td>
<td>152</td>
<td>14.81%</td>
</tr>
<tr>
<td></td>
<td>Somewhat religious</td>
<td>268</td>
<td>26.12%</td>
</tr>
<tr>
<td></td>
<td>Not very religious</td>
<td>147</td>
<td>14.33%</td>
</tr>
<tr>
<td></td>
<td>Not at all</td>
<td>441</td>
<td>42.98%</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>18</td>
<td>1.75%</td>
</tr>
<tr>
<td>Education</td>
<td>Below high school graduate</td>
<td>4</td>
<td>0.39%</td>
</tr>
<tr>
<td></td>
<td>High school graduate</td>
<td>102</td>
<td>9.94%</td>
</tr>
<tr>
<td></td>
<td>Some college or Associate's degree</td>
<td>236</td>
<td>23.00%</td>
</tr>
<tr>
<td></td>
<td>Bachelor's degree</td>
<td>450</td>
<td>43.86%</td>
</tr>
<tr>
<td></td>
<td>Graduate or professional degree or higher</td>
<td>134</td>
<td>13.06%</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>10</td>
<td>0.97%</td>
</tr>
<tr>
<td>Income</td>
<td>Less than $20,000</td>
<td>170</td>
<td>16.57%</td>
</tr>
<tr>
<td></td>
<td>$20,000 to $34,999</td>
<td>210</td>
<td>20.47%</td>
</tr>
<tr>
<td></td>
<td>$35,000 to $49,999</td>
<td>217</td>
<td>21.15%</td>
</tr>
<tr>
<td></td>
<td>$50,000 to $74,999</td>
<td>231</td>
<td>22.51%</td>
</tr>
<tr>
<td></td>
<td>$75,000 to $99,999</td>
<td>114</td>
<td>11.11%</td>
</tr>
<tr>
<td></td>
<td>$100,000 to $149,999</td>
<td>45</td>
<td>4.39%</td>
</tr>
<tr>
<td></td>
<td>$150,000 or more</td>
<td>15</td>
<td>1.46%</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>24</td>
<td>2.34%</td>
</tr>
</tbody>
</table>

Table 1. Demographic information of the 1,026 MTurk survey participants.
Discussion

Most of the research on the impact of genetic testing to date has taken place in the context of clinical care or medical research and has demonstrated that people are often motivated by health care concerns and have a variety of worries about how the data are used.\textsuperscript{23, 24} The views of individuals regarding DTC-GT,\textsuperscript{25} which by definition is not mediated through the health care system and which is discretionary, provide additional perspectives on the public’s opinions about genetic testing. One critical observation is that, despite increasing advertisement of these services and the growing numbers of people who have enrolled,\textsuperscript{26} these highly computer-literate individuals largely have not pursued DTC-GT and do not know many people who have. The latter fact is particularly important since they report that the views of relatives and close friends would be more important in shaping their interest than social media. This observation underscores the need to learn more about the impact of social factors on the uptake of genetic testing, influences that to date have been under examined.

These respondents stated that they would be most interested in obtaining ancestry information from DTC-GT, which is consistent with the fact that genealogy and ancestry are the second most common hobby and reason for internet searches in the United States.\textsuperscript{27} They ultimately would be driven to use these services, however, to address their own and their relatives’ health issues. This is concerning given the limited amount of FDA-authorized genetic health reports currently on the market\textsuperscript{4} and early indications that much of this information may be of questionable utility to consumers.\textsuperscript{28, 29}

Like many others, these respondents overwhelmingly seek to control their data and express a variety of concerns about who would get access to their data. Given the wide variation in the data privacy policies adopted by DTC-GT companies\textsuperscript{8} and the efforts of some companies to monetize their consumers’ genetic data, these respondents’ desires to have control and to be protected are not likely to be met. At the same time, these respondents were quite willing to permit data to be used for research, especially if they were de-identified, as has been observed in many previous studies.\textsuperscript{23, 30} Their greater preference to allow data about them to be used by academic researchers than by companies, the government, and people outside the United States is concordant with what has been observed in earlier studies.\textsuperscript{31, 32} Respondents also expressed varying degrees of concern about law enforcement access to DTC-GT test results in the context of criminal investigations. Respondents overwhelmingly believed that law enforcement should have very little access to DTC-GT data in context of crime prevention, but were much more permissive of law enforcement use of such data to solve crimes.

These members of Amazon Mechanical Turk, like many research participants, generally would want to obtain additional health-related results as they became available.\textsuperscript{33} Yet only a quarter say that feel confident that they can interpret these data without assistance, with more than 50% saying that they would appreciate expert assistance and the remainder saying that they would require that help. Since DTC-GT companies generally do not provide counseling, these results will place additional burdens on a health care system that already has insufficient numbers of geneticists and genetic counselors to meet the needs of patients who need clinical care. Indeed, reports indicate that many consumers are beginning to seek guidance on DTC-GT results from their primary care physicians, who may also be ill-equipped to interpret the results. Furthermore, emerging evidence suggests that returning results may, in some cases, lead to health care interventions of unclear value.\textsuperscript{34}

Limitations

Although a small cohort of undergraduate students at Vanderbilt University was able to correctly interpret our survey questions, this does not rigorously demonstrate that the broader MTurk population found all questions thoroughly understandable. MTurk has been shown to offer a diverse population that includes significant ranges of age, education level, and socioeconomic status. But while demographically diverse in some respects, we note that MTurk workers tend to be predominately white, and are generally younger and more highly educated when compared to the entire U.S. population (See Table 1).\textsuperscript{35} Potential distinctions within our survey population remain largely unexamined. For example, non-white respondents may not share similar concerns with white respondents relating to law enforcement’s involvement with DTC-GT. Furthermore, young respondents may generally have less experience regarding the social pressures inherent in having children. It is also worth noting that because the majority of our respondents reported that they had not undergone DTC-GT, most questions were hypothetical for them. Nonetheless, our results provided intriguing signals about areas that require further study.
Conclusion

DTC-GT is a source of individual genetic information that is particularly fraught with incomplete and often misleading results. Despite concerns about the validity and utility of DTC-GT, it appears as though at-home genetic testing is here to stay for a host of purposes. In particular, as consumers attempt to decipher their health risks using such services, increasing numbers of people will turn to clinicians, seeking understanding of their results and guidance about managing their health. This has led to increasing debate about how providers and DTC-GT companies should respond to this influx.

The American College of Medical Genetics and Genomics (ACMG) claimed that DTC-GT companies should be more transparent by indicating their laboratory’s accreditation, clearly communicating privacy policies, demonstrating the validity and utility of their testing options, and providing consumers with access to genetics experts. It is evident that there is a growing demand for reliable interpretation of DTC-GT results, which neither providers nor DTC-GT are prepared to fulfill. While it would be beneficial to minimize the inappropriate use of healthcare resources based on DTC-GT results, it is crucial to educate providers about DTC-GT so that they may help patients who have concerns about their testing results. The medical informatics community should take charge of this issue by developing decision support for clinicians to handle such situations.

Acknowledgements

This work was sponsored in part by the Center for Genetic Privacy and Identity in Community Settings (GetPreCiSe) through a grant from the National Human Genome Research Institute, National Institutes of Health (#RM1HG009034). The authors wish to thank the members of the center for their assistance in the review of the survey questions and refinement of the study.

References


782
Clinical Tractor: A Framework for Automatic Natural Language Understanding of Clinical Practice Guidelines

Daniel R. Schlegel, PhD1, Kate Gordon1, Carmelo Gaudioso, MD, PhD2, Mor Peleg, PhD3
1SUNY Oswego, Oswego, NY, USA; 2Roswell Park Cancer Institute, Buffalo, NY, USA; 3University of Haifa, Haifa, Israel

Abstract  Computational representations of the semantic knowledge embedded within clinical practice guidelines (CPGs) may be a significant aid in creating computer interpretable guidelines (CIGs). Formalizing plain text CPGs into CIGs manually is a laborious and burdensome task, even using CIG tools and languages designed to improve the process. Natural language understanding (NLU) systems perform automated reading comprehension, parsing text and using reasoning to convert syntactic information from unstructured text into semantic information. Influenced by successful systems used in other domains, we present the architecture for a system which uses NLU approaches to create semantic representations of entire CPGs. In the future, these representations may be used to generate CIGs.

1 Introduction
Clinical practice guidelines (CPGs) — documents designed to assist in diagnosis and treatment of disease, developed by professional organizations through systematic review of the literature, and usually distributed in a natural language (e.g., English) — form the foundation of evidence-based medicine. It is well known that compliance with paper guidelines is lacking, but that compliance improves greatly with the introduction of clinical decision support systems (CDSS), which implement guideline recommendations and are integrated into electronic health record (EHR) systems (see, e.g., [1]). Were the semantic content of the CPGs represented computationally as Computer Interpretable Guidelines (CIGs) the task of building CDSS would be eased.

Over the past two decades, methods and formalisms have been developed for representing guidelines computationally as CIGs (see [2]), but there are still few actively maintained CIGs since the process of creating them is extremely time-consuming and burdensome. If the creation of CIGs were a one-time effort perhaps the burden of manual curation could be overcome, but guidelines change frequently (often annually) in complex ways which would require ongoing effort. Therefore we believe that effort should be placed on automatically generating CIGs from their paper counterparts. In order to do so accurately and comprehensively, we believe it is necessary to represent the semantic content of CPGs to the greatest extent possible. This semantic representation may then be used to create the associated CIG. Here we present a framework and proof of concept based on natural language understanding (NLU) techniques to automatically represent the semantic content of the CPG. We leave rigorous evaluation and CIG generation from the semantics for future work.

Natural language understanding is a subtopic of natural language processing in which the goal is to build a computer system which performs reading comprehension on a given input text. These techniques are currently not widely used in the biomedical informatics community in part because the language used is complex, presupposing a significant amount of implicit knowledge. There is also a need for high precision due to the safety-critical domain. Implementing custom tools to perform the NLU task while addressing these issues requires wide-ranging expertise (biomedicine, computational linguistics, and knowledge representation and reasoning) and can be labor intensive.

The framework presented here adopts its high-level design from a previous NLU system, Tractor, designed for understanding short intelligence messages in the counter-insurgency (COIN) domain [3, 4, 5], and adapts it to the clinical domain. The Tractor system was successful in its task — it converted input text to a knowledge base (KB) containing over 92% semantic relations using rules that fired correctly nearly 98% of the time [4]. In internal evaluation not yet published we found the transformation to be on par with what a human is capable of performing. Our new system, currently under active development, is dubbed Clinical Tractor.

The remainder of the paper is organized as follows. First we aim to convince the reader that adopting portions of the Tractor architecture is appropriate while also showing where the differences lie. Clinical Tractor’s architecture is detailed in Section 3 along with a worked example in Section 4. As it is helpful to have an understanding of Clinical
Tractor when considering related research on language processing with CPGs, we save the discussion of related work for (Section 5), where we also discuss the future of the Clinical Tractor project.

2 A Comparison of the Domains of Tractor and Clinical Tractor

Tractor was initially developed for the COIN domain, requiring a large portion of reality to be modeled. Persons about whom intelligence messages are written are usually performing the activities of daily life – shopping, driving, making phone calls, interacting with other persons, carrying items, etc. The persons and items are described in varying amounts of detail. Problematically, it is unknown in advance which of these activities or attributes will be important when the messages are combined to form a complete picture of what is happening in an area. This uncertainty forced Tractor to be developed in a highly general way, so as to model a large number of activities and attributes at once, modeling specifics only where the general models were insufficient.

In this regard the domain of clinical medicine is significantly simpler. In general there is only a single person being discussed, the patient (though discussions of family history may also be present). In guidelines there is another person, the clinician, who is asked to perform some actions. The attributes of import and the actions that are or should be taken encompass only those related to health, not all of reality. A significant advantage to working in the clinical domain is the existence of a wide variety of controlled vocabularies, terminologies, and ontologies, which allow the identification of a large number of these actions and attributes. The strategy to model these actions and attributes can be adopted from Tractor with little modification. In fact, the general rules used for modeling many activities and attributes at once can be used with little or no modification.

Clinical guidelines have the advantage of containing, in general, grammatical text. Intelligence messages, like medical records, do not share this property – they often contain sentence fragments, semi-structured components, and unconventional punctuation/abbreviations. The Tractor system was built to be somewhat resilient to these issues, using only surface features where possible, working around mistakes made by the linguistic parsers in non-grammatical portions of text, and containing a system for specifying the components of semi-structured text. With CPGs we do not anticipate significant issues of this form, except perhaps in inclusion criteria where it is present, but it will be significant in planned future extensions of the work to include EHR data. Guidelines do contain some structured components in the form of document structure, which we account for.

Whereas intelligence messages are a record of what has happened, CPGs suggest what is to happen in the form of recommendations. This is significant as recommendations in CPGs often contain modal verbs, qualifying action phrases with words such as “should”, “may”, “might consider”, and so forth, in general covering the modalities of likelihood, ability, permission, and obligation. These, importantly, provide a weight to the recommendation. Weights also may be derived from the degree of evidence upon which the recommendation is based, usually provided on a scale somewhere in the guideline.

Intelligence messages and individual CPG recommendations are both similar in that they are short, avoiding issues such as topic shift and rhetorical/discourse relations. On the other hand, CPG recommendations often have temporal semantics dictated by their order. Within sections of the narrative of a guideline topic shift is generally avoided, and some discourse relations that arise in storytelling are eschewed. There is the potential for sections to exhibit rhetorical relations such as narrative strengthening [6], though we do not believe this requires any architectural additions.

3 Clinical Tractor

Because of the domain differences, the architecture of Clinical Tractor is different from that of Tractor, with more of a focus on extraction of data based on document structure, and making use of background knowledge. The architecture, seen in Figure 1, consists of four main components: text processing using various processing resources (PRs) operating within the open-source General Architecture for Text Engineering (GATE) [7]; converting the GATE output to a syntactic KB consisting of propositions in a first order logic; aligning terms in the KB with background knowledge and importing relevant data; and mapping syntactic relations to semantic relations using both domain-neutral and domain-specific mapping rules informed by the background knowledge.¹

¹We outline the framework making use of tools we have selected, but these tools could be swapped for others which perform the same tasks.
3.1 Input Data

Guidelines are distributed in several formats. In order to standardize them for our pipeline, we manually convert the guidelines of interest to an XML format capturing the document structure such as headings, tables, and inset boxes. We also include graph structures in figures and algorithms (as in the NCCN guidelines). No guideline-specific semantic features are included. The XML format used [8] is based on a combination of the Journal Article Tag Suite (JATS) [9] and GraphML [10]. In the future this transformation would either be an automated process, or, in a more ideal future, guidelines would be distributed in (possibly one of several) standardized format(s).

3.2 Text Processing in GATE

Each input CPG is processed by a set of PRs operating within GATE. Most of these PRs are from the ANNIE (a Nearly-New Information Extraction System) suite [7]. Shown in Figure 1 are: the ANNIE English Tokenizer and Sentence Splitter that divide the input into linguistic units; the Stanford Dependency Parser, for part-of-speech tagging and parsing (discussed further in Section 3.2.1); the GATE Morphological Analyser for identifying root forms of inflected nouns and verbs; a group of named-entity recognizers – list based, ontology-based, rule based, and MetaMap (discussed further in Section 3.2.2); and a group of PRs that perform co-reference resolution. GATE uses a plugin architecture allowing for the use of many other PRs as well as the creation of custom PRs. It also allows customization of each of the selected PRs according to the domain.

3.2.1 Dependency Parsing

The notion of dependency relations in language is ancient, going back to Pāṇini’s grammar in the 5th century BCE. Phrase structure grammar which is more commonly covered in introductory linguistics courses, on the other hand, is a modern invention. Dependency grammars represent syntactic structure as (often binary) relations between tokens in the text. These relations are known as dependencies. The Universal Dependencies [11] used in the Stanford Dependency Parser contain mostly syntactic relations, but also relations consistent with a shallow semantic parse. This apparently semantic information makes the task of developing syntax-semantics mapping rules to determine semantic roles somewhat easier.

---

2 There are purely algorithmic transformations from phrase structure parses to dependency parse (e.g., [12]), meaning there isn’t inherently more semantic information present in the dependency parse – the information is simply organized in a more useful way.
### 3.2.2 Named Entity Recognition

Named entity recognition (NER) is a component often associated with information extraction (IE) systems in which structured data is extracted from free text for one or more classes of entities. These classes often include the names of persons, locations, and organizations, but also dates, addresses, quantities, etc. As discussed above, the person entities in a guideline are fairly straightforward. In contrast, there are very large classes of entities such as drugs, procedures, diseases, symptoms, and anatomical locations. There are also a significant number of entities related to measurements and temporal relation. Guidelines also include evidence levels represented in various forms. NER can be used to identify the text spans representing an instance of a class of entities, and can also assist in unifying the the multiple ways of expressing a concept (e.g., anatomical location) in natural language. We make use of several forms of NER:

- **List-Based NER** GATE contains a “gazetteer” PR for identifying entities from lists. Lists may contain complete named entities such as names or locations, or words (keys) which given context can indicate that a named entity begins with or ends with the key (e.g., “Hospital” in a hospital name or “Jr.” in a person’s name). A key type of great importance in CPGs is that which indicates the current sentence, paragraph, or section is or contains recommendations. Gazetteer items have a major and minor type, allowing for a shallow ontology.

- **Ontology-Based NER** Related to list-based NER is ontology based NER. Terms and their synonyms are identified in the text through simple matching. We have developed tools [13] which extract this data from ontologies to store them in gazetteer lists for matching. By storing them in lists, additional synonyms can easily be added without modifying the underlying ontology.

- **MetaMap** Probably the most popular method for recognizing terms from medical vocabularies in text, and making use of the UMLS, MetaMap is sometimes criticized for its precision/recall. In combination with other approaches, it can be a useful addition to a complete NER suite.

- **Rule-Based NER** Rules allow identification of named entities through regular expressions over annotations using the Java Annotation Patterns Engine (JAPE). These rules allow for recognition of complete entities for which keys were noted in the list-based NERs. Entities with semi-structured formats such as prescription drugs may also be recognized. Rules provide an opportunity for a first pass at disambiguation and the removal of over-matches given the context available in word orderings.

Downstream processing in the syntax-semantics mapping rules makes use of the dependency parse to perform the bulk of the NLU task. One source of confusion in designing an NLU system of this type is how much NER to do using rules and lists, and how much to do later using the dependency parse. Dependency parsing captures structural relationships (dependencies) well, but recognition based on word order given a dependency parse is quite difficult. Therefore we limit ourselves to recognition which is word-order dependent at this stage.

### 3.3 Propositionalizer

The result of GATE processing is a set of annotations, each consisting of an identifier, a start and end position within the CPG’s text, a type, and a set of attribute-value pairs. Each of GATE’s PRs produces these annotations, so the set consists of information about XML document structure, tokens, sentences, paragraphs, dependencies, named entities, etc. The propositionalizer converts the set of GATE annotations into a set of logical propositions.

Given the input from GATE, the propositionalizer merges annotations which have the same start and end positions (e.g., a token and one or more results from the NERs). The result of this is a set of annotations each with unique start and end positions, and each with a unique identifier. The propositionalizer re-constructs the hierarchy of document-related XML tags and produces logical assertions in a form subsuming that of DoCO, the Document Components Ontology [14]. In addition to what DoCO offers, head words of sentences (found via the dependency parse) are attached to the sentences for use in the syntax-semantics mapper.

The propositionalizer produces a KB consisting of a set of propositions (expressions which may have a truth value assigned to them), in the logical language of the CSNePS knowledge representation and reasoning (KRR) system [15, 16, 17]. CSNePS is used to represent and perform reasoning on all of the KBs created by Clinical Tractor from the English CPGs. CSNePS is simultaneously a logic-, frame-, and graph-based KRR system [18]. It is the latest member of the SNePS family of KRR systems [19].
A CSNePS proposition may be “asserted” meaning it is taken to be true in the KB. Propositions need not be asserted to exist in the KB; CSNePS can consider propositions of unknown truth. When we discuss “asserting a proposition” we mean to add it to the KB as an assertion, and when we discuss “unasserting a proposition” we mean to remove the assertion from the KB. CSNePS uses a term logic, in which all expressions are terms – even those that in first order logic would not be. This means that propositions may have propositions as arguments (allowing for meta-knowledge). This is especially useful for representing the source of knowledge. The arguments of a proposition are terms that could denote words, tokens, syntactic categories, entities and events, and classes or properties of these entities and events.

Relations and the propositions in which they occur may be categorized as either: syntactic, taking as arguments terms denoting words, tokens, and syntactic categories; or as semantic, taking as arguments entities and events in the domain and their classes and properties. A KB is syntactic to the extent that its assertions are syntactic, and is semantic to the extent that its assertions are semantic. The KB produced by the propositionalizer is mostly syntactic, and therefore is referred to as the syntactic KB.

3.4 Background Knowledge Alignment

The syntactic KB is enhanced by a background knowledge alignment system (BKAS). This system matches spans of text against lexical resources such as WordNet and VerbNet, and locates ontological terms based on the results of NER in SNOMED (via MetaMap CUIs) and biomedical ontologies. The matched data is imported into the KB. Where the data is hierarchical as in the WordNet hypernym hierarchy, the VerbNet hierarchy, and ontological subsumption hierarchies, relevant hierarchies are imported into the KB. Where other logical relations are present, those are imported as well. Background knowledge allows mapping rules to be written more generally – for example, instead of operating only on a specific verb or list of verbs, a rule might operate on classes of related verbs by using a higher level concept.

Clinical Tractor is designed to operate in an ontologically heterogeneous environment, in which a single concept in the text may be annotated with multiple ontological terms from different sources. That said, we make heavy use of the OBO Library ontologies which have been co-developed to be inter-operable, as the more entirely separate sources there are, the more complex downstream processing in the mapping rules becomes.

The BKAS is meant to be generic, allowing for the easy addition of resources as needed. In future work, we intend for this to include knowledge extracted from other materials such as journal articles and other guidelines. As the BKAS system enhances the syntactic KB with background knowledge, we refer to the result as an enhanced syntactic KB.

3.5 Syntax → Semantics Mapper

The enhanced syntactic KB is operated on by mapping rules, converting the mostly syntactic KB to a mostly semantic representation. Whereas IE approaches aim to identify “within text instances of specified classes of entities and of predications involving these entities” [20, emphasis added], we aim to convert the entire syntactic content of the guideline into semantic content, doing true automatic reading comprehension. This includes understanding all parts of the text, not only verb relations or noun phrases matching some pre-specified patterns as other systems do (see Section 5). The mapping rules are represented in the CSNePS rule language and are executed within the CSNePS KR system.

The mapping rules, designed to be generic, come in two major types – those that convert syntactic representations to more easily processable syntactic representations, and those that convert syntactic representations to semantic ones. The left side of Figure 2 shows a rule that simplifies syntactic representations, transforming phrases in the passive voice to the active voice. This rule fires (i.e., is executed) when an nsubjpass (passive nominal subject) relation is identified in the dependency parse. This relation occurs between a verb and its passive subject. It converts this into a dobj (direct object) relation and unasserts the nsubjpass relation. The rule also looks to see, in a subrule, if the verb is in a case relation with the word “by”, and makes the nominal subject of the verb the object of the prepositional relation. This rule would transform the parse of “morphine should be prescribed by the clinician” to the parse of “the clinician should prescribe morphine”. In building NLU systems, the number of rules can quickly grow out of hand; rules such as this simplify the problem somewhat by requiring no special rules to be written for handling passive phrases.
The right side of Figure 2 shows two syntax-semantics mapping rules. The first of these, dobjAction, would make morphine the theme of the ‘prescribe’ action in the previous example. The relation theme reflects one of the linguistic thematic relations [21, 22, 23], often used to express the action of a verb. This rule fires when the verb, prescribe in this case, is a member of the class Action. This is derivable from background knowledge sources. The dobjPerception rule fires when the verb is a Perception action, a more specific case than dobjAction. A verb would be known to be a perception by making use of imported data from the BKAS, such as VerbNet. In this case the topic thematic role is used. For example, a guideline might contain the text “... when complications are discovered.” Here complication is the topic of discover. Determination of the thematic roles to use is done by making use of the Unified Verb Index [24] wherever possible.

Background knowledge sources play an important role in the mapping rules. The lexical relations available from WordNet and VerbNet allow the creation of general rules which are specific to the kinds of things discussed in guidelines. While an above example makes use of the “prescribe” verb, many others could be used (e.g., receive, take, be given). In general, these verbs have some medication or treatment as their direct object, and indicate a transference of ownership. Verbs of this type are covered by a small set of upper-level concepts in the lexical resources, which may be used in the mapping rules. While our goal is to use general rules wherever possible, we will use more specific rules as discussed here when necessary. Using this technique, the rules can identify many of the “Action Palette” [25] action types used in guidelines.

As Clinical Tractor is still under development, the principal effort is in building suitable mapping rules. Development involves the creation of new domain-neutral rules and many more domain-specific rules. We aim to understand noun phrases, including those that otherwise might be given only a single code by an NER system. Using a common, consistent, semantic structure exposes the relation between long expressions which may not have a code, such as “cellulitis of left hallux” and shorter expressions (e.g., “cellulitis”) and other long expressions that do have codes, such as “cellulitis of toe of left foot”. Negation may also be understood using mapping rules.

The mapping rules also must handle verb phrases, these include condition-action phrases, the Action Palette items, and discussions of choices and decision making. These often include evidence in the form of in-text citations, statements of evidence level, and modality. Guidelines also provide plans for treatment, or give guidance on creating such plans which must be handled. The result of applying the mapping rules is a semantic KB.

4 A Worked Example

Consider the following recommendation from the ADA Standards of Medical Care in Diabetes 2017 [26]: “Patients found to have elevated blood pressure should have blood pressure confirmed on a separate day.” To illustrate the pipeline, this was processed through an early prototype of the system with limited NER (using only MetaMap), only
MetaMap Tagging (GATE): Patients found to have elevated blood pressure should have blood pressure confirmed on a separate day.

Dependency Parsing

CSNePS Syntactic Graph for “Patients found to have elevated blood pressure”

CSNePS Semantic Graph after mapping rules applied

Figure 3: A small example using a prototype implementation of Clinical Tractor, with some by-hand augmentation. MetaMap matches and the dependency parse for a single recommendation are shown at the top, with a subset of the CSNePS syntactic KB after propositionalization, and semantic KB after the mapping rules have been applied shown for the phrase “Patients found to have elevated blood pressure.”

The processing results are shown in Figure 3. Identified concepts in MetaMap are shown on top, with the Stanford dependency parse directly underneath. The propositionalizer converted the GATE output to a CSNePS KB, and we’ve visualized it as a propositional graph. Portions of the KB for the text “patients found to have elevated blood pressure” are shown. Each token can be seen attached to its identifier (beginning with n). Dependency relations may be seen in the graph. The string representation of the multi-word expression “elevated blood pressure” and its decomposition into single words can be seen at the bottom right of the syntactic graph. MetaMap CUIs and concept names for some of the tokens are also shown. This graph excludes many additional relations that are in the KB for easier readability.

The result of applying the mapping rules is shown at the bottom of the figure in the semantic graph. A subtle change that has occurred is that n terms that originally denoted syntactic entities now denote semantic entities. Previously n40 denoted a token with the text “patients”. Now it denotes a group of entities, each of which is of the type patient. n17 was a token with the text “pressure”, adjectivally modified by “elevated” but now denotes an entity of type pressure, with the modifiers elevated and blood. The MetaMap concept Increase in blood pressure applies to this entity instead of the string “elevated blood pressure” as it did in the syntactic graph. This entity is possessed by n40. In sum, this graph represents the group of patients possessing elevated blood pressure.

Only two changes to the semantic graph were handled manually for this example. Two MetaMap concepts were identified for “elevated blood pressure” - a disorder (hypertensive disease), and a finding (increase in blood pressure). We manually selected the finding concept, though the process for selecting the correct one is already well defined: “found” is past tense of “find”, a member of the verb frame for “discover” (using VerbNet) which indicates a clinical observation; a finding. Second, we moved MetaMap concepts for multi-word expressions to the head noun.
Clinical Tractor shares characteristics with many systems used in biomedical informatics for language processing tasks. Guideline-focused work tends to be centered on the task of aiding the creation of CIGs by performing IE tasks to retrieve, and possibly restructure, salient portions of the CPG. Several examples make use of semantically informed patterns over the text. Wenzina and Kaiser [27] use patterns over UMLS semantic types to identify condition-action sentences. They observed recall of 75% and precision of 88% on a small evaluation set. In other work, Kaiser, et al. worked to identify treatment activities in guideline text [28]. Here they used the UMLS semantic network types and relations to generate semantic patterns for activities such as performing, using, and analyzing. They made use of lists of verbs corresponding to the relations and a dependency parse to determine which MetaMap identified concepts in the sentence fit the the subject and object of the relation.

Serban and colleagues [29] presented an ontology-driven method for pattern matching on frequently recurring linguistic patterns, mapped to the control structures (e.g., sequencing, if-then, action-effect) of the target CIG formalism. Medical thesauri have been used [30] to enhance the ability to identify portions of guideline text which map to reusable building blocks, useful for guideline formalization.

Machine learning techniques have also been applied to the recommendation identification task. Preliminary work extracting regular-expression-based heuristic patterns has shown some promise [31], but the inclusion of semantic data is needed. Other work has used part of speech tags as features to extract action sentences from CPGs [32], but again, without the use of semantic data. Neither of these approaches were specifically tailored to the CPG domain. While not using machine learning approaches, Taboada et al. [33] provide evidence for the need to tailor systems to the domain at hand. They used several off-the-shelf tools to extract descriptive knowledge about therapeutic and diagnostic procedures, finding that adaptation of the tools to the task improved results, though their paper doesn’t make tailoring vs. non-tailoring directly comparable.

It’s important to note that in isolation each of the above systems cover only a small subset of what is necessary to derive the complete semantics of the recommendations of a CPG, let alone an entire CPG. In considering the approach taken by Clinical Tractor, there are similarities to the pattern-matching approaches. The primary difference is that instead of using UMLS resources combined with mostly surface structure, Clinical Tractor aims to make significant use of the dependency parse, other kinds of NER, additional background knowledge resources to enhance generalization, and rules applied in multiple steps to move toward a completely semantic representation. This goal of a complete semantic representation appears to be unique to Clinical Tractor in the domain of CPGs.

In relying on, at least in part, a more “standard” NLP pipeline, Clinical Tractor also shares some characteristics with systems such as cTAKES [34], CLAMP [35], HITEx [36], and HTP-NLP [37]. The domain in which these tools are most commonly used is slightly different: information extraction for electronic health records (EHRs). While these domains have many similarities such as including some narrative structure, mentioning many of the same kinds of named entities, and being patient-centric, there are differences in content and appearance of the language. Medical records tend to contain more non-standard abbreviations, sentence fragments, and non-standard English. While only occasionally appearing in guidelines themselves, clinical protocol eligibility criteria do often share the property of being made up of sentence fragments. Work on ERGO [38] discusses many of the challenges involved in eligibility criteria, including acronyms, Boolean operators, and comparison statements. We have in-progress work addressing these issues mostly during the rule-based NER stage of Clinical Tractor. Even given the challenges of this type of text, we don’t believe there’s a need to go as far as to use IE components which are meant explicitly to account for such problems, such as NegEx, which appear frequently in EHR-focused pipelines using the above tools.

6 Conclusion

We have presented an architecture for an NLU system meant to perform, as near as possible, complete reading comprehension of CPGs. The early stages of our architecture use a fairly standard NLP pipeline, comparable to what many NLP systems in biomedicine perform, though with enhanced NER capabilities. It is unique in the attention paid to aligning background knowledge with the textual contents and going a step further than pattern matching rules over text, making use of a KRR system and syntax-semantics mapping rules to transform a syntactic KB into a semantic one utilizing NLU techniques. This approach is built upon that taken by Tractor, which has shown to be successful in
Clinical Tractor is currently under active development as part of a larger system for the automatic generation of CIGs from CPGs. Our hope is that this systems proves useful to people working in and researching biomedicine, and that over time we can build a compendium of semantically represented knowledge. Our even longer term goal for Clinical Tractor is to generalize it to work wherever there is text in biomedicine, whether it be in guidelines, EHRs, journal articles, or clinical trial protocols. Moreover it is important to us that what we build be free for the world to use; as components of our system reach a usable state they will be released open source, under a non-restrictive license.

Acknowledgements

This work was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R15LM013030. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


Using Natural Language Processing to improve EHR Structured Data-based Surgical Site Infection Surveillance

Jianlin Shi, MD MS1, Siru Liu, BS3, Liese C.C. Pruitt, MD1, Carolyn L. Luppens, MD1, Jeffrey P. Ferraro, PhD1,2, Adi V. Gundlapalli, MD, PhD1,3, Wendy W. Chapman, PhD1, Brian T. Bucher, MD MS1

1 School of Medicine, University of Utah, Salt Lake City, Utah, US
2 Intermountain Healthcare, Salt Lake City, Utah, US
3 VA Salt Lake City Healthcare System, IDEAS Center 2.0, Salt Lake City, Utah, US

Abstract

Surgical Site Infection surveillance in healthcare systems is labor intensive and plagued by underreporting as current methodology relies heavily on manual chart review. The rapid adoption of electronic health records (EHRs) has the potential to allow the secondary use of EHR data for quality surveillance programs. This study aims to investigate the effectiveness of integrating natural language processing (NLP) outputs with structured EHR data to build machine learning models for SSI identification using real-world clinical data. We examined a set of models using structured data with and without NLP document-level, mention-level, and keyword features. The top-performing model was based on a Random Forest classifier enhanced with NLP document-level features achieving a 0.58 sensitivity, 0.97 specificity, 0.54 PPV, 0.98 NPV, and 0.52 F0.5 score. We further interrogated the feature contributions, analyzed the errors, and discussed future directions.

Introduction

Healthcare-associated infections (HAIs) affect 10% of patients undergoing surgical procedures in the United States and contribute a significant amount of morbidity to patients1. For patients, the development of an HAI can be associated with increased risk of mortality, longer length of stay, and increased risk of readmission2. Given the significance of HAIs to healthcare utilization, several surgical quality improvement programs have been designed to aid in the surveillance of HAIs. These programs include the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) and the Centers for Disease Control (CDC) National Healthcare Safety Network3. The backbone of these programs is the ascertainment of clinical data through a manual chart review processes4. These processes, while providing high-quality information, are inherently limited by the cost and effort of chart review5. Additionally, these processes are under high risk of underreporting HAIs6, 7, caused by biased interpretation of guideline or data, use of different criteria, insufficient training, etc. To account for these limitations, the programs rely on a sampling methodology either by selecting specific procedures or by a random sampling of procedures, thus limiting the scalability and generalizability of the manual chart review process.

Surgical site infections (SSI) are the most common type of HAI occurring after surgical procedures2. According to the CDC, an SSI is an infection in the surgical wound occurring within 30 days of the operative procedure and can be classified as superficial, deep, or organ space SSI depending on the location of the infection with the surgical wound8. One important distinction for classification of SSIs is to identify the cases with an infection present at the time of surgery (PATOS) as PATOS infections significantly increase the risk of a recurrence of an SSI after the procedure. The diagnosis of SSI can occur through multiple pathways, including provider diagnosis, provider-initiated treatment, or isolation of organisms from a microbiology culture. Given the heterogeneous nature of diagnosing SSIs, manual chart review is currently required to verify or exclude diagnoses9.

Over the past ten years, in response to the Health Information Technology for Economic and Clinical Health Act, there has been rapid adoption of Electronic Health Records (EHR) in US hospitals9. As of 2017, 96% of federal acute care hospitals have adopted a Certified EHR capable of computerized decision support, computerized provider order entry and electronic physician documentation. The almost universal adoption of certified EHRs has to potential to leverage the secondary use of EHR data for quality assurance programs and reduce or depreciate the effort of the manual chart review for surveillance of SSIs.

In previous work, several groups have leveraged electronic health care data to create automated surveillance systems
for SSIs. Hu et al. used EHR data from a single institution along with machine learning models to detect the occurrence of HAIs after surgical procedures. They used International Classification of Diseases (ICD) Codes, laboratory results, microbiology orders and results, medication administration, radiology orders, and vital signs to develop logistic regression models, using manually chart reviewed data as the gold standard. Overall, the area under the curve (AUC) for the classification of SSI was 0.91. The most predictive features included microbiology results and imaging orders. The limitation of this study is the lack of clinical text information in the development of the classification models. Given that the diagnosis of SSI is mainly dependent on physician diagnosis that is usually described in clinical notes, the inclusion of text information may improve the performance of the classification model.

Grundmeier et al. expanded upon this work to include limited text information to build classification models for SSI. Pre-specified keywords related to SSI that had been identified from clinical narratives were added using regular expression matching. The overall performance of the classification model significantly improved with the addition of keywords (AUC 0.97 vs. 0.94). Their best model performed with a 0.9 sensitivity and 0.28 positive predictive value (PPV). The most informative keyword features they found included: “pain,” “infection,” “warm,” “drainage.” While the inclusion of keywords improves the model performance, ignoring the context of keywords is a significant limitation of this approach, which we elaborated further in the discussion.

Other groups have focused exclusively on natural language processing methodology. FitzHenry et al. developed a rules-based NLP system using SNOMED-CT concepts from free text notes in the Veterans Affairs Healthcare system. They developed rules for classifying ten postoperative complications after surgical procedures, including SSI. The performance of the system for detection of SSI was modest, with a sensitivity of 77% and specificity of 63%. The advantage of this approach is the inclusion of text-based features for SSI classification. However, they did not include other structured data into their algorithms, such as medications and laboratory values.

In the present study, we aim to build upon previous work on automated detection of SSIs using EHR data. We combine machine learning techniques with a rules-based natural language processing system. We hypothesize the inclusion of NLP-derived features in structured data-based machine learning models will improve the classification performance for detection of SSIs after surgical procedures. To aim at real-world use, we target our cohort to all surgical patients without any complication code filtering.

Methods

In this study, we applied a rule-based natural language processing (NLP) to extract the information related to surgical site infections (SSI) from the clinical notes, and infer the infection status for each document. Then we combined the keywords or the NLP derived features with structured EHR data to build machine learning models to classify the patients’ SSI status.

Machine learning data set development

The patient data was collected through the University of Utah Health Enterprise Data Warehouse (EDW), including patients from the University Hospital and Clinics, the University of Utah from July 1st, 2013 to August 31st, 2017. The inclusion criteria were patients whose surgical course underwent review by a trained surgical clinical reviewer (SCR) as part of the ACS-NSQIP program. Patients were excluded if they expired within 30 days of the operation. For patients included in the present study, we obtained the following structured EHR data from their records between 2-30 days after their operative procedure: hospital encounters, ICD 9th/10th Edition Diagnosis Codes, medication administration, radiology orders, laboratory and microbiology results, and vital signs. A total of 5,795 patients were included, among which 291 patients were labeled as SSI positive by manual chart review performed by an SCR. For the text data, we created a full corpus by obtaining all clinical notes from 0-30 days after the operative procedure of the following note types: admission notes, history and physical notes, discharge notes and progress notes. We excluded notes that were incomplete or not signed by a credentialed provider (Physician or Advanced Practice Provider).

Structured EHR Data Feature Engineering

For each category of features from the structured EHR data, we performed feature engineering for the development of machine learning models. ICD codes were mapped to the Agency for Healthcare Research and Quality Clinical Classification software and treated as a binary variable. Medications were mapped to RxNorm, and the number of
days of each medication class was used as features. For radiology orders, we grouped common orders and used the number of orders for that particular group as a feature. Laboratory results and vital signs were grouped based on clinical meanings. For example, Potassium levels from both serum or plasma were grouped. Due to the temporal nature of laboratory and vital sign measurement, we used only the minimum, median, and maximum values of each group as features. Groups with greater than 25% missing elements were excluded. For groups with <25% missing values, we imputed using the median of the group. For the microbiology results, we created binary features based on the culture type and organism isolated.

Development of NLP Annotation Guidelines

Based on the ACS NSQIP definitions and clinical expertise, we developed an annotation guideline to perform iterative manual annotation of the NLP development corpus. We defined a mention-level annotation type \textit{Evidence of SSI} and a document-level annotation type \textit{SSI Status}. The \textit{Evidence of SSI} was used to annotate the sentences that explicitly or implicitly mentioned any evidence of infection of a surgical wound. Infection mentions without explicitly referring to an incision were not annotated. For instance, we ignored “he doesn’t have signs of infection.” The \textit{Evidence of SSI} had two attributes: \textit{Infection Status} and \textit{Temporality}. The \textit{Infection Status} was used to clarify if an SSI statement was affirmative, probable, or negated. \textit{Temporality} was used to distinguish current and past SSIs. The document-level annotation type \textit{SSI Status} has one attribute \textit{Infection Status} with the same three eligible values above: affirmative, probable, or negated. At this step, we did not distinguish between SSI and SSI PATOS.

Creation of Gold-Standard NLP-Development Corpus

We created an NLP-development corpus for manual annotation from a random sample of 78 patients from the ACS NSQIP flagged SSI positive cases. The notes were selected using the same criteria mentioned above and cut off at total 2000 notes. We elected to only include SSI positive cases in our NLP development corpus due to the low prevalence of SSIs. Sampling data from positive patients improved the efficiency of both annotation and NLP rule development. In addition, because SSIs develop over the postoperative course, clinical notes from patients with SSIs contain a combination of SSI positive and negative documents.

These clinical notes were pre-annotated using Knowledge Author\textsuperscript{13} with broad vocabularies consolidated based on CDC and NSQIP definitions and using vocabulary expansion from the Unified Medical Language System (UMLS). The corpus was then split into 50 document batches and imported into the annotation tool: eHOST (The Extensible Human Oracle Suite of Tools)\textsuperscript{14}. Two surgeons with expertise in SSI (LCCP, CLL) were trained using the annotation guidelines and evaluated using inter-annotator agreement. After training on five batches, the inter-annotator agreement reached over 80\% for mention level and over 90\% for document-level annotation. A third reviewer (BTB) checked the disagreements and performed adjudication. We subsequently performed a single-person annotation on the remaining 1750 documents with 10\% overlap between the annotators to ensure consistent agreement. On the overlapped notes the annotators had an inter-annotator agreement of 70\% at the mention level and 85\% at the document level. The annotated gold-standard NLP-development corpus was split 50\% for NLP pipeline training (1000 documents) and testing (1000 documents) corpus.

NLP Pipeline Development and Evaluation

The NLP pipeline was developed using a rule-based tool, Easy Clinical Information Extractor (EasyCIE)\textsuperscript{15}. The main components of EasyCIE are listed in Table 1. These rule-based components were built on an optimized rule processing engine\textsuperscript{16} to allow high-speed processing of large rule sets and data sets. The rules for section detector, sentence segmenter, and negation context detector were imported from previous projects with minor tuning. The rest of the rules were developed based on the 1,000 NLP pipeline training documents. The developed pipeline was assessed on the training set for its rule coverage and then evaluated against the NLP pipeline test corpus (1000 notes). Because the rule-based NLP system used more granular and additional mention types to infer document-level conclusions, its mention-level annotation types did not match with the reference standard. Thus, we did not evaluate the mention-level performance, and only document-level F1 scores are reported.

Machine Learning Model Development and Evaluation

After the NLP system was developed from the NLP corpus, we processed all the text data in the machine learning
Table 1: NLP components and corresponding functionality of EasyCIE

<table>
<thead>
<tr>
<th>NLP Components</th>
<th>Functionality Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section detector</td>
<td>Identify the sections, e.g., History of Present Illness, Family History</td>
</tr>
<tr>
<td>Sentence Segmenter</td>
<td>Detect sentence boundaries</td>
</tr>
<tr>
<td>Named entity recognizer</td>
<td>Identify target concepts using dictionaries</td>
</tr>
<tr>
<td>Context detector</td>
<td>Attach the context information as feature values to the corresponding target concepts</td>
</tr>
<tr>
<td>Attribute inferencer</td>
<td>Make mention-level conclusions based on target concepts and corresponding attributes</td>
</tr>
<tr>
<td>Document inferencer</td>
<td>Make a document-level conclusion from the corresponding mention-level conclusions</td>
</tr>
</tbody>
</table>

corpus to develop NLP features. These features included two binary feature sets: an NLP document feature set (ND) and an NLP mention feature set (NM). The document feature set included two features: \( ND_{SSI\ POS} \) to cover whether a patient’s documents contain any positive SSI note; and \( ND_{SSI\ PROB} \) to represent whether a probable SSI note was included. The mention feature set includes 28 features corresponding to the 28 mention-level annotation types generated by NLP. For instance, \( NM_{ABDOMEN} \) represents the feature describing if there is any mention that can indicate the abdomen body site; \( NM_{INCISION} \) represents the feature describing if any incision mention exists.

In addition, two numeric feature sets were also created from NLP document-level results. Frequency (\( FRE \)): the total number of the positive SSI notes or the possible SSI notes of a patient during the observation time window. Normalized frequency (\( NOR \)): the \( FRE \) divided by the total notes of that patient during the observation time window. In total, we obtained six document-level features. We performed the same conversion to mention-level annotations, resulting in 84 mention-level features. As another baseline comparison, we simply used the named entity recognizer without any context detection and inferences to extract the keywords in the notes. The extracted mentions were used to generated keyword features (\( KW \)). For readers’ convenience, we listed the abbreviations of different feature sets in Table 2 below.

Table 2: Abbreviations of the feature sets used in this study

<table>
<thead>
<tr>
<th>Abbreviations*</th>
<th>Feature level</th>
<th>Feature type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO_NLP</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ND_BIN</td>
<td>Document</td>
<td>Binary</td>
</tr>
<tr>
<td>ND_FRE</td>
<td>Document</td>
<td>Frequency</td>
</tr>
<tr>
<td>ND_NOR</td>
<td>Document</td>
<td>Normalized frequency</td>
</tr>
<tr>
<td>ND_ALL</td>
<td>Document</td>
<td>All three types above</td>
</tr>
<tr>
<td>NM_BIN</td>
<td>Mention</td>
<td>Binary</td>
</tr>
<tr>
<td>NM_FRE</td>
<td>Mention</td>
<td>Frequency</td>
</tr>
<tr>
<td>NM_NOR</td>
<td>Mention</td>
<td>Normalized frequency</td>
</tr>
<tr>
<td>NM_ALL</td>
<td>Mention</td>
<td>All three types above</td>
</tr>
<tr>
<td>KW_BIN</td>
<td>Keywords</td>
<td>Binary</td>
</tr>
<tr>
<td>KW_FRE</td>
<td>Keywords</td>
<td>Frequency</td>
</tr>
<tr>
<td>KW_NOR</td>
<td>Keywords</td>
<td>Normalized frequency</td>
</tr>
<tr>
<td>KW_ALL</td>
<td>Keywords</td>
<td>All three types above</td>
</tr>
</tbody>
</table>

Note: Each of these abbreviations represents a set of features instead of a single feature. *: All the above feature sets include structured data features.

We performed a Chi-Square feature selection to select the top performing variables. Logistic regression, support vector machines (SVM), and random forest (RF) models were developed. We applied nested cross-validation to the data set to avoid overfitting17. Because the data set was highly imbalanced with relatively sufficient positive cases, we performed down-sampling rather than up-sampling to avoid overfitting. We used three folds in the outer loop and five folds in the inner loop. We used a grid search for parameters in the inner loop. The best estimators from the inner loop were selected based on \( F_1 \) score because the data used for inner loop cross-validation were down sampled. Then these best estimators were evaluated on test sets of the outer loop (without down-sampling). We used the following performance metrics for evaluation: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), plus \( F_{0.5} \). We choose the F measure with \( \beta =0.5 \), because we prefer to weight precision higher than recall for
practical surveillance consideration. Due to the imbalanced data set, we elected not to use the area under the receiver operating curve (AUC) in the assessment. We compared the sensitivity and specificity of each model to the NO_NLP baseline machine learning model using McNemar’s test. In addition, we used bootstrapping with 1000 replications to obtain the distribution of $F_{0.5}$ score and compared the distribution of each model to the NO_NLP baseline machine learning model.

**Results**

Once the rule-based NLP solution was developed, we assessed the rule coverage by calculating the performance on the training set. Our final micro-average of precision, recall, and F1 scores were 0.83, 0.83, 0.83 on the training corpus, respectively. Then we evaluated the NLP system on the 1000 blind test notes. The micro-average of the precision, recall, and F1 scores were 0.75, 0.75, 0.75 in the testing corpus, respectively. Next, we used the developed NLP solution to process all the text data of our full corpus to generate NLP features for machine learning. To demonstrate the value of the NLP-derived features, Figure 1 plots the frequency of selected NLP-derived features in patients with and without SSI. There are significantly more SSI-related NLP features in SSI positive compared to SSI negative patients ($p<0.01$).

![Figure 1: The means and standard deviations of example frequency features within SSI positive and negative cohorts](image)

We subsequently used the NLP-derived features to develop the machine learning models for patient-level classification of SSI. We evaluated models using the features described in Table 2 and compared the performance of each model to the baseline structured data model without NLP-derived features. The results are shown in Table 3. The RF models with document features have the highest average Specificity (0.97) and significantly improved compared to the baseline RF NO_NLP model ($p<0.001$). The best PPV (0.54) and best $F_{0.5}$ (0.52) comes from the RF models enhanced with binary document features ($ND_{BIN}$). Compared with the RF models with structured data features alone ($NO_{NLP}$), the PPV significantly improved by 8% and $F_{0.5}$ improved by 7% ($p<0.001$). The addition of mention-level features or keyword features did not significantly improve the $F_{0.5}$ compared the baseline RF model with structured data alone. Although the logistic regression models and the SVM models have higher NPV values (0.99), their $F_{0.5}$ scores are lower than the corresponding RF models, because they are lower in PPV. The SVM models with binary mention features ($NM_{BIN}$), all mention features ($NM_{All}$) and all NLP features ($ND_{NM}$) also have the highest sensitivity, but much worse $F_{0.5}$. The highest $F_{0.5}$ of all the models enhanced with keywords features is 0.480, which is not significantly different from the models with structured data alone.

**Discussion**

*The value of NLP in SSI detection*

In previously reported studies of surgical site infection surveillance, structured variables (e.g., ICD Codes, laboratory results, microbiology orders and results, medication administration, etc.) are commonly used to generate machine
Table 3: Machine learning models’ performances (average scores) on the test sets without balancing

<table>
<thead>
<tr>
<th>Models</th>
<th>Feature sets*</th>
<th>Sensitivity (p-value)#</th>
<th>Specificity (p-value)#</th>
<th>PPV</th>
<th>NPV</th>
<th>F0.5 (p-value)#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>NO_NLP</td>
<td>0.70</td>
<td>0.95</td>
<td>0.46</td>
<td>0.98</td>
<td>0.47</td>
</tr>
<tr>
<td>Forest</td>
<td>ND_BIN</td>
<td>0.58 (&lt;0.001)</td>
<td>0.97 (&lt;0.001)</td>
<td>0.54</td>
<td>0.98</td>
<td>0.54 (&lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>ND_FRE</td>
<td>0.59 (&lt;0.001)</td>
<td>0.97 (&lt;0.001)</td>
<td>0.53</td>
<td>0.98</td>
<td>0.53 (&lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>ND_NOR</td>
<td>0.59 (&lt;0.001)</td>
<td>0.97 (&lt;0.001)</td>
<td>0.53</td>
<td>0.98</td>
<td>0.52 (&lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>ND_ALL</td>
<td>0.57 (&lt;0.001)</td>
<td>0.97 (&lt;0.001)</td>
<td>0.53</td>
<td>0.98</td>
<td>0.52 (&lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>NM_BIN</td>
<td>0.59 (&lt;0.001)</td>
<td>0.96 (&lt;0.001)</td>
<td>0.51</td>
<td>0.98</td>
<td>0.50 (0.99)</td>
</tr>
<tr>
<td></td>
<td>NM_FRE</td>
<td>0.70 (0.56)</td>
<td>0.94 (0.02)</td>
<td>0.45</td>
<td>0.98</td>
<td>0.46 (0.08)</td>
</tr>
<tr>
<td></td>
<td>NM_NOR</td>
<td>0.69 (0.65)</td>
<td>0.95 (0.81)</td>
<td>0.46</td>
<td>0.98</td>
<td>0.47 (0.93)</td>
</tr>
<tr>
<td></td>
<td>NM_ALL</td>
<td>0.58 (&lt;0.001)</td>
<td>0.96 (&lt;0.001)</td>
<td>0.51</td>
<td>0.98</td>
<td>0.50 (0.73)</td>
</tr>
<tr>
<td></td>
<td>ND_NM</td>
<td>0.59 (&lt;0.001)</td>
<td>0.96 (&lt;0.001)</td>
<td>0.52</td>
<td>0.98</td>
<td>0.51 (0.14)</td>
</tr>
<tr>
<td></td>
<td>KW_BIN</td>
<td>0.58 (&lt;0.001)</td>
<td>0.96 (&lt;0.001)</td>
<td>0.51</td>
<td>0.98</td>
<td>0.50 (0.63)</td>
</tr>
<tr>
<td></td>
<td>KW_FRE</td>
<td>0.70 (0.32)</td>
<td>0.95 (0.29)</td>
<td>0.45</td>
<td>0.98</td>
<td>0.47 (0.57)</td>
</tr>
<tr>
<td></td>
<td>KW_NOR</td>
<td>0.60 (&lt;0.001)</td>
<td>0.96 (&lt;0.001)</td>
<td>0.52</td>
<td>0.98</td>
<td>0.51 (0.06)</td>
</tr>
<tr>
<td></td>
<td>KW_ALL</td>
<td>0.59 (&lt;0.001)</td>
<td>0.96 (&lt;0.001)</td>
<td>0.51</td>
<td>0.98</td>
<td>0.50 (0.84)</td>
</tr>
<tr>
<td>SVM</td>
<td>NO_NLP</td>
<td>0.80</td>
<td>0.91</td>
<td>0.32</td>
<td>0.99</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>ND_BIN</td>
<td>0.80 (0.48)</td>
<td>0.91 (0.61)</td>
<td>0.33</td>
<td>0.99</td>
<td>0.37 (0.63)</td>
</tr>
<tr>
<td></td>
<td>ND_FRE</td>
<td>0.79 (0.65)</td>
<td>0.91 (0.04)</td>
<td>0.33</td>
<td>0.99</td>
<td>0.38 (0.11)</td>
</tr>
<tr>
<td></td>
<td>ND_NOR</td>
<td>0.80 (1.00)</td>
<td>0.91 (0.09)</td>
<td>0.32</td>
<td>0.99</td>
<td>0.36 (1.00)</td>
</tr>
<tr>
<td></td>
<td>ND_ALL</td>
<td>0.80 (1.00)</td>
<td>0.91 (0.68)</td>
<td>0.32</td>
<td>0.99</td>
<td>0.37 (0.01)</td>
</tr>
<tr>
<td></td>
<td>NM_BIN</td>
<td>0.81 (0.53)</td>
<td>0.91 (0.60)</td>
<td>0.33</td>
<td>0.99</td>
<td>0.37 (0.01)</td>
</tr>
<tr>
<td></td>
<td>NM_FRE</td>
<td>0.78 (0.06)</td>
<td>0.91 (0.10)</td>
<td>0.33</td>
<td>0.99</td>
<td>0.37 (0.15)</td>
</tr>
<tr>
<td></td>
<td>NM_NOR</td>
<td>0.80 (1.00)</td>
<td>0.92 (&lt;0.001)</td>
<td>0.34</td>
<td>0.99</td>
<td>0.38 (1.00)</td>
</tr>
<tr>
<td></td>
<td>NM_ALL</td>
<td>0.81 (0.39)</td>
<td>0.92 (0.16)</td>
<td>0.34</td>
<td>0.99</td>
<td>0.38 (0.01)</td>
</tr>
<tr>
<td></td>
<td>ND_NM</td>
<td>0.81 (0.49)</td>
<td>0.91 (0.21)</td>
<td>0.34</td>
<td>0.99</td>
<td>0.38 (0.01)</td>
</tr>
<tr>
<td></td>
<td>KW_BIN</td>
<td>0.80 (0.81)</td>
<td>0.91 (0.48)</td>
<td>0.32</td>
<td>0.99</td>
<td>0.36 (0.62)</td>
</tr>
<tr>
<td></td>
<td>KW_FRE</td>
<td>0.79 (0.26)</td>
<td>0.92 (&lt;0.001)</td>
<td>0.35</td>
<td>0.99</td>
<td>0.39 (0.01)</td>
</tr>
<tr>
<td></td>
<td>KW_NOR</td>
<td>0.79 (0.26)</td>
<td>0.91 (0.05)</td>
<td>0.33</td>
<td>0.99</td>
<td>0.38 (0.34)</td>
</tr>
<tr>
<td></td>
<td>KW_ALL</td>
<td>0.80 (1.00)</td>
<td>0.91 (0.86)</td>
<td>0.32</td>
<td>0.99</td>
<td>0.36 (0.86)</td>
</tr>
<tr>
<td>Logistic</td>
<td>NO_NLP</td>
<td>0.77</td>
<td>0.95</td>
<td>0.43</td>
<td>0.99</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>ND_BIN</td>
<td>0.78 (0.18)</td>
<td>0.95 (1.00)</td>
<td>0.44</td>
<td>0.99</td>
<td>0.48 (0.92)</td>
</tr>
<tr>
<td></td>
<td>ND_FRE</td>
<td>0.77 (0.32)</td>
<td>0.95 (0.04)</td>
<td>0.45</td>
<td>0.99</td>
<td>0.49 (0.15)</td>
</tr>
<tr>
<td></td>
<td>ND_NOR</td>
<td>0.77 (1.00)</td>
<td>0.95 (1.00)</td>
<td>0.43</td>
<td>0.99</td>
<td>0.48 (0.15)</td>
</tr>
<tr>
<td></td>
<td>ND_ALL</td>
<td>0.79 (0.10)</td>
<td>0.94 (&lt;0.001)</td>
<td>0.40</td>
<td>0.99</td>
<td>0.45 (0.70)</td>
</tr>
<tr>
<td></td>
<td>NM_BIN</td>
<td>0.79 (0.18)</td>
<td>0.94 (&lt;0.001)</td>
<td>0.40</td>
<td>0.99</td>
<td>0.44 (0.47)</td>
</tr>
<tr>
<td></td>
<td>NM_FRE</td>
<td>0.78 (0.65)</td>
<td>0.95 (&lt;0.001)</td>
<td>0.45</td>
<td>0.99</td>
<td>0.49 (0.60)</td>
</tr>
<tr>
<td></td>
<td>NM_NOR</td>
<td>0.77 (1.00)</td>
<td>0.95 (1.00)</td>
<td>0.43</td>
<td>0.99</td>
<td>0.48 (0.01)</td>
</tr>
<tr>
<td></td>
<td>NM_ALL</td>
<td>0.79 (0.18)</td>
<td>0.94 (&lt;0.001)</td>
<td>0.40</td>
<td>0.99</td>
<td>0.44 (0.11)</td>
</tr>
<tr>
<td></td>
<td>ND_NM</td>
<td>0.79 (0.11)</td>
<td>0.94 (&lt;0.001)</td>
<td>0.40</td>
<td>0.99</td>
<td>0.44 (0.17)</td>
</tr>
<tr>
<td></td>
<td>KW_BIN</td>
<td>0.79 (0.17)</td>
<td>0.93 (&lt;0.001)</td>
<td>0.39</td>
<td>0.99</td>
<td>0.43 (&lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>KW_FRE</td>
<td>0.77 (1.00)</td>
<td>0.95 (1.00)</td>
<td>0.43</td>
<td>0.99</td>
<td>0.48 (1.00)</td>
</tr>
<tr>
<td></td>
<td>KW_NOR</td>
<td>0.77 (1.00)</td>
<td>0.95 (1.00)</td>
<td>0.43</td>
<td>0.99</td>
<td>0.48 (1.00)</td>
</tr>
<tr>
<td></td>
<td>KW_ALL</td>
<td>0.79 (0.17)</td>
<td>0.93 (&lt;0.001)</td>
<td>0.39</td>
<td>0.99</td>
<td>0.43 (&lt;0.001)</td>
</tr>
</tbody>
</table>

*: The abbreviations of feature sets are explained in Table 2, NO_NLP, structured data features alone.
ND, NLP document-level features; NM, NLP mention-level features; BIN, binary features;
FRE, frequency features; NOR, normalized frequency features; ALL, use all document or mention features.
#: Computed against the models using structured data alone.
learning features for disease classification. These variables can reflect a significant amount of information related to SSIs, including diagnostic information as well as treatment modalities for SSI. However, given the complex nature of SSI diagnosis, without the inclusion of textual information, these models are inherently limited in performance. For example, the documentation of SSI may only be located in the clinical notes if the infection did not result in any change of laboratory orders, vital sign or medication administration, e.g., a mild surgical wound infection. Also, laboratory results or medication administration may not exist in the structured database if a patient has received emergent care outside the index hospital. This type of missing information varies a lot across different health organizations; it is especially prevalent in tertiary medical centers where patients commonly travel a long distance to receive care.

Previous studies have investigated incorporating free text into machine learning models using keywords; however, without contextual information the use of keywords does not maximize the potential of NLP. In this study, we observed that it is common that clinical notes mentioned the keywords related to SSI but not in the way of demonstrating evidence of an SSI. For example, surgical procedures in the past medical history or any medical conditions mentioned as patient education for future precautions. Of all the mentions (980,045 mentions) that we extracted using our NLP systems, 7.3% are negated concepts (e.g., “Midline incision well approximated with staples, no erythema, no drainage”), 1.5% are within irrelevant context (e.g., “Stay away from people with infections”), 3.5% are about other infections other than abdominal SSIs (e.g., “septic arthritis of the bilateral wrists”), and 0.5% are historical concepts (e.g., “Sepsis 05/2011”). Ignoring the context of these mentions adds in a significant amount of noise (12.8% in total) to machine learning models. Additionally, a document-level conclusion of an SSI may not stem from the explicit mention of SSI, but instead from inferencing multiple mentions of other different concepts. For instance, a note with mentions of “post-op s/p” and “intraabdominal infection” in separated sentences. Neither of these two concepts alone can conclude SSI, while together, they suggest a positive SSI. These cases require additional inference from mention-level information to draw document-level conclusions, whereas the simple keywords approach is incapable of.

The results of the comparative experiments above (Table 3 suggest that handling these subtle contexts and logic correctly does provide additional information gains in the development of machine learning models. The best F0.5 score using structured data alone is 0.48 (Logistic regression model). Adding keywords features improves the performance on the RF models compared with RF models using structured data alone, but the best F0.5 of all keywords enhanced models is still 0.48. Using NLP features, the best F0.5 score reaches 0.52 – an 8.3% improvement (p<0.05).

NLP error analyses and indications for future directions

The NLP errors occur at two levels: mention level, document level. The most common mention-level errors are from the context not seen in the training set. For instance, “site color red” is mentioned in a list of description of a surgical incision. Our NLP solution did not capture the earlier context about the incision, thus failed to detect the “red” as a sign of infection. In the document-level errors, the most common cause is that our developed rules do not capture all the logic at the document level. For instance, we did not observe the rare co-occurrence of other healthcare-associated infection, such as pneumonia or urinary tract infection (UTI) in the NLP training set. In these cases, it is common to refer to pneumonia or UTI as “infection” in general. Within a context of abdominal condition description, NLP tends to interpret these general infection mentions as possible abdomen infections. However, in the NLP test set, these cases were more prevalent. Among all the 180 false negative SSI notes, 35 of them have a diagnosis of pneumonia, and 64 notes have other infections without an explicit mention of an anatomical site. These are the errors that we will address in future studies.

Additionally, we also found some annotation errors in the reference standard. Often these errors happened within the notes where doctors use templates or copy a section of text from other notes, while changing a few words that support different document conclusions. The annotators tended to skip the snippets that they have seen before, and these slightly modified text pieces become difficult to identify. On the other hands, it suggests that pre-annotating these notes using developed NLP solutions might improve the human review accuracy and reduce the workload. Implementing these into real practice workflow could speed up the human chart review process, reduce errors and human labor, even if a completely automated surveillance system is not available.

Machine learning models analyses and indications for future directions
As the results show, NLP features do provide additional value to the machine learning models. Notably, our champion RF models using NLP document features not only scored better than the others but also have statistically significant differences compared with the models using structured data alone.

The logistic regression models generated sub-optimal results among the three classifiers. We examined the best performing models built from NLP document frequency features (ND_FRE). Because we used nested cross-validation, we trained 15 logistic regression models enhanced with the ND_FRE feature set (5 models in each outer loop). Then we got three best models (one per each outer loop). We identified the common features among these three models, by using the average of the corresponding coefficients and ranked them in descending order (see Table 4). The results rank the NLP derived document-level feature — the frequency of SSI positive documents, as the No. 1 weighted feature.

**Table 4:** The average coefficients of the common features used in the three best Logistical Regression models

<table>
<thead>
<tr>
<th>Feature names</th>
<th>Average coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND_SSI_POS_F</td>
<td>5.21</td>
</tr>
<tr>
<td>Anaerobic_Culture</td>
<td>3.74</td>
</tr>
<tr>
<td>Abscess</td>
<td>2.97</td>
</tr>
<tr>
<td>Guided</td>
<td>2.68</td>
</tr>
<tr>
<td>CCS_Dx_Postoperative_Infection</td>
<td>2.56</td>
</tr>
<tr>
<td>Wound_Culture</td>
<td>1.83</td>
</tr>
<tr>
<td>CCS_Dx_Peritonitis_&amp;_intestinal_abcess</td>
<td>0.87</td>
</tr>
<tr>
<td>CCS_Dx_Comlications_of_surgical_procedures_or_medical_care</td>
<td>0.62</td>
</tr>
</tbody>
</table>

We also examined the Random Forest classifier by estimating the importance of features (See Table 5). It shows that the NLP document feature is ranked as the most important feature. We chose one of the decision trees within the RF classifier to visualize its tree structure. Figure 2 visualizes a decision tree’s root from one of the top RF models, which identified the NLP binary document feature stands as the tree root. 56.7% of the test cases are classified as positive, and 43.3% are classified as negative. Moreover, 80% of the cases on both sides are correctly classified. Other trees show similar functions to these features. These visualizations indicate that NLP does play an essential role in our champion RF models. Because of the overall performance of SVM models (they are the lowest among the three classifiers in this study), and the difficult nature of feature interpretation for SVM, we do not present the SVM models analyses here.

**Table 5:** The feature permutation importance from the best RF models using NLP document binary features

<table>
<thead>
<tr>
<th>Feature names</th>
<th>Feature importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND_SSI_POS</td>
<td>0.20</td>
</tr>
<tr>
<td>Med_DIAGNOSTIC_PRODUCTS</td>
<td>0.19</td>
</tr>
<tr>
<td>CCS_Dx_Postoperative_Infection</td>
<td>0.16</td>
</tr>
<tr>
<td>CT_abd_pelvis</td>
<td>0.13</td>
</tr>
<tr>
<td>Anaerobic_Culture</td>
<td>0.11</td>
</tr>
<tr>
<td>Wound_Culture</td>
<td>0.09</td>
</tr>
<tr>
<td>CCS_Dx_Peritonitis_and_intestinal_abcess</td>
<td>0.04</td>
</tr>
<tr>
<td>CCS_Dx_Peritonitis</td>
<td>0.03</td>
</tr>
<tr>
<td>CCS_Dx_Comlications_of_surgical_procedures_or_medical_care</td>
<td>0.02</td>
</tr>
<tr>
<td>Guided</td>
<td>0.02</td>
</tr>
<tr>
<td>Abscess</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Additionally, we investigated the errors predicted by machine learning models and NLP. We found a significant number of errors are caused by misclassifying the patients who admitted with infections present at the time of surgery (PATOS), e.g., perforated appendix or cholecystitis. Among these patients, some patient went on to develop infectious complications after surgery and are labeled as SSI PATOS. Others do not develop infectious complications and are not labeled as SSI or SSI PATOS. Our structured data features neither include any temporal information such as when a
Figure 2: Visualizing the root of a decision tree from one of the best Random Forest models enhanced with document binary features. “samples” means the percentage of cases reaching the node. “value”: left is the percentage of true negative cases, right is the percentage of true positive cases. “class” is the conclusion from the parent node.

lab was ordered, lab values were collected, nor anti-infection treatments were administrated. Also, our current NLP system does not discriminate the infection status at the time of surgery or later in the postoperative course. Thus our machine learning models and NLP system have difficulty correctly classifying these patients. Since these patients are not rare in general surgery, we will address the issue in future studies. This barrier might be overcome either by integrating temporal features into machine learning models or adding higher level inferencer of NLP from document-level conclusions to patient level conclusions.

Compared with previously published studies\textsuperscript{10,12}, our methods are inherently different. The reported scores are not directly comparable. FitzHenry et al. used manipulated input data (patients with complications mixed with 10% of the ones without complications) to develop and evaluate their system. Thus the reported performance does not reflect the system’s real performance on real-world data. Hu et al. compared several different models using data without any manipulation, but only reported AUC. Because both of our data are highly skewed with negative cases, only a small portion of ROC AUC can effectively reflect true performance. For this reason, we did not report ROC AUC. Hu et al. did not report the exact AUC scores for SSI detection other than a plot. Our several models (e.g., SVM enhanced with document binary features, Logistic regression with all document features or mention features) reach 0.92 in AUC, which seems comparable to their best model. However, they are not the best model that we choose using $F_{0.5}$ score. Because we built and evaluated our models on a real-world data set, these models can be more practically used in the real environment.

Conclusion

We assessed the value of applying NLP to enhance machine learning models for SSI detection. The results show improved performance of the NLP enhanced models compared with the models using structured data alone or models enhanced with keywords features. A close investigation of these enhanced models confirms that NLP features play an important role in classification. Our error analysis suggests that future studies should include temporal status into consideration to further improve the performance.

Acknowledgments

We acknowledge funding and resources: the Agency for Healthcare Research and Quality, 1K08HS025776 (BTB) and the VA Salt Lake City Center of Innovation Award (COIN) #I50HX001240 from the US Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Health Services Research and Development (AVG). The views expressed in this article are those of the authors and do not necessarily represent the position or policy of the US Department of Veterans Affairs or the US government.

References


2. R P Merkow, M H Ju, J W Chung, B L Hall, M E Cohen, M V Williams, T C Tsai, C Y Ko, and K Y Bilimoria. Underlying reasons associated with hospital readmission following surgery in the United States. JAMA,


Have ICD-10 Coding Practices Changed Since 2015?

Srikanth Sivashankaran, MS, John P. Borsi, MS, Amanda Yoho, PhD
IBM Watson Health, Cleveland, OH

Abstract

Usage of ICD-10 codes in administrative data has continued to shift since mandatory adoption in 2015. Identifying changing patterns in coding behavior is imperative in producing reliable analyses and robust conclusions. We examined the granularity of ICD-10 coding over time in a cohort selected from the IBM Explorys Therapeutic Dataset, which contains the records of over 60 million patients. Our seasonality-aware trend model identified patterns of interest, such as increased use of laterality codes for pain and increased use of codes denoting concepts novel to ICD-10 for screening encounters. Those relying on these codes should adjust for these ‘learning curve’ effects. This work should be extended to additional modalities of terminology usage and represents a starting point for researchers working with dynamic clinical ontologies.

Introduction

The Tenth Revision of the International Classification of Diseases (ICD-10) represented a “thorough rethinking” of the World Health Organization (WHO) terminology. Specifically, the clinical modification of ICD-10 (ICD-10-CM) in use in the United States included a four-fold increase in the number of available codes over the ICD-9 version, intended to dramatically increase the applicability, specificity, and usability of the code system. The migration to ICD-10 in the US proved controversial: on the one hand, the conversion to ICD-10 promised improved understanding of patient outcomes, reduced fraud and abuse, and increased extensibility. On the other, the increase in complexity and structural changes threatened to incur implementation costs and cause headaches for practitioners.

We should expect the disruption surrounding the 2015 transition to have an observable impact on the resulting ICD-coded data. A 2001 report from the CDC highlighted the statistical disruption produced by ICD-10 coding for mortality data, and reviews from the 1998-99 Australian and 2002-07 Canadian ICD-10 transitions showed noticeable changes in coding volume and prevalence of some disease categories. And, indeed, US studies in the wake of the cutover have shown under- and over-reporting of important clinical categories. Substance abuse, self-harm, HIV and Alzheimer’s coding were subject to significant alterations in perceived population prevalence across the ICD-10 boundary.

The impact of the conversion to ICD-10 is not limited to the immediate transition period, however. In the wake of Switzerland’s 1998 ICD cutover, for example, ICD usage continued to shift for years as coders acclimated to the new system. The idea of a coding ‘learning curve’ is supported both by anecdotal and experimental evidence. A South African controlled study demonstrated that ICD coding behavior can noticeably shift as a result of training and familiarization. Early reviews of US ICD coding post-cutoff show that we might expect a significant learning curve; as of 2016 there remained significant opportunities to improve the accuracy and relevance of coding.

Although learning curve effects may contribute substantially to coding variation immediately following a transition, there are other factors that might impact coding patterns over time. Coders may respond to risk adjusted payment incentives or different auditing regimes. There are potentially any number of systemic factors that could have subtle influences on subsequent data analyses. Understanding and revalidating ICD-10 coding patterns is an important step towards having confidence in analyses using ICD-coded data moving forward.

We provide a unified structure for assessing temporal coding patterns over the years 2015-2018 and present results from applying it to a large administrative data set. We examine both seasonal and long-run trends for related families of codes. For this work, we focus primarily on codes which were newly introduced in ICD-10, to best capture the relevant shifts in usage following the ICD-10 transition period.

Data and Methods

We employed the General Equivalence Mappings (GEMs) from the Center for Medicare and Medicaid Services (CMS) to identify novel concepts in ICD-10-CM. These bidirectional mappings, between ICD-9-CM and ICD-10, are accompanied by a series of flags, one of which indicates whether the mapping between concepts in the two
terminologies is approximate. We identified novel ICD-10 codes by applying all of the following conditions in the ICD-10 to ICD-9 segment of the diagnosis terminology GEMs:

1. The “approximate” flag had the value TRUE
2. The ICD-10 term did not describe a “not elsewhere classified” or “other” concept
3. No other ICD-9 term mapped to the ICD-10 term under consideration
4. The ICD-9 term in the GEM entry describes an “other” concept – a catch-all for terms not explicitly enumerated in that terminology

These criteria identify, for example, the ICD-10 code for Salmonella pyelonephritis (A02.25) as describing a concept not listed in ICD-9:

<table>
<thead>
<tr>
<th>ICD-10 Code</th>
<th>ICD-10 Description</th>
<th>Approximate mapping?</th>
<th>ICD-9 Code</th>
<th>ICD-9 Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A02.25</td>
<td>Salmonella pyelonephritis</td>
<td>TRUE</td>
<td>003.29</td>
<td>Other localized salmonella infections</td>
</tr>
</tbody>
</table>

In order to provide a baseline against which to compare the use of novel codes, we organized the concept space of ICD-10 into groups consisting of codes sharing a header. Headers in ICD-10 are grouping artifacts defined by CMS as “not valid” but “included as a convenience for other uses”, to capture related sets of terms. For this study, we considered every concept group meeting the following conditions:

1. The group contained an ICD-10 term identified in the previous step
2. The membership of the group remained constant over the study period (fiscal years 2016, 2017, and 2018 to align with the ICD code update cycle)
3. The group contained at least one term describing an “other”, “unspecified”, or “not elsewhere classified” concept – one to be used where the medical record lacks adequate granularity or describes a condition not otherwise enumerated -- against which to compare the novel ICD-10 term(s)

These criteria situate the novel ICD-10 concept of Salmonella pyelonephritis, for example, within the following concept group:

<table>
<thead>
<tr>
<th>Header</th>
<th>Header Description</th>
<th>Code</th>
<th>Code Description</th>
<th>Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>A02.2</td>
<td>Localized salmonella infections</td>
<td>A02.20</td>
<td>Localized salmonella infection, unspecified</td>
<td>(unspecified)</td>
</tr>
<tr>
<td>A02.2</td>
<td>Localized salmonella infections</td>
<td>A02.21</td>
<td>Salmonella meningitis</td>
<td>(non-novel)</td>
</tr>
<tr>
<td>A02.2</td>
<td>Localized salmonella infections</td>
<td>A02.22</td>
<td>Salmonella pneumonia</td>
<td>(non-novel)</td>
</tr>
<tr>
<td>A02.2</td>
<td>Localized salmonella infections</td>
<td>A02.23</td>
<td>Salmonella arthritis</td>
<td>(non-novel)</td>
</tr>
<tr>
<td>A02.2</td>
<td>Localized salmonella infections</td>
<td>A02.24</td>
<td>Salmonella osteomyelitis</td>
<td>(non-novel)</td>
</tr>
<tr>
<td>A02.2</td>
<td>Localized salmonella infections</td>
<td>A02.25</td>
<td>Salmonella pyelonephritis</td>
<td>(novel)</td>
</tr>
<tr>
<td>A02.2</td>
<td>Localized salmonella infections</td>
<td>A02.25</td>
<td>Salmonella with other localized infection</td>
<td>(other)</td>
</tr>
</tbody>
</table>

We then obtained unique patient counts per month for each ICD-10 code in the selected groups over the fiscal years 2016, 2017, and 2018 from the IBM Explorys Therapeutic Dataset. These data are comprised of the clinical and claims records of over 60 million unique patients from over 39 healthcare systems; see Table 1 for demographic information about the study cohort selected from those patients. In order to more precisely consider the behavior of professional medical coders, we limited the study to administrative data alone. To ensure sufficient sample size, we limited analyses to the 20 header groups with greatest overall patient counts.
Table 1. Cohort demographic summary. IQR stands for interquartile range.

<table>
<thead>
<tr>
<th>Total Individuals [Count]</th>
<th>7,893,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex [Count (%)]</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>4,486,000 (56.8%)</td>
</tr>
<tr>
<td>Male</td>
<td>3,406,000 (43.1%)</td>
</tr>
<tr>
<td>Other/unidentified</td>
<td>2,000 (0.1%)</td>
</tr>
<tr>
<td>Race [Count (%)]</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>5,878,000 (74.5%)</td>
</tr>
<tr>
<td>Black</td>
<td>803,000 (10.2%)</td>
</tr>
<tr>
<td>Asian</td>
<td>214,000 (2.7%)</td>
</tr>
<tr>
<td>Other/unidentified</td>
<td>729,000 (9.2%)</td>
</tr>
<tr>
<td>Ethnicity [Count (%)]</td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>6,139,000 (77.8%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>626,000 (7.9%)</td>
</tr>
<tr>
<td>Declined</td>
<td>394,000 (5.0%)</td>
</tr>
<tr>
<td>Other/unidentified</td>
<td>734,000 (9.3%)</td>
</tr>
</tbody>
</table>

In order to surface patterns in the assignment of ICD-10 codes over time, we measured the difference between novel codes and other codes. We decomposed the observed variation in differences into level, trend, and seasonality components using the R forecast package, which selects an exponential smoothing method for each component that minimizes the value of Akaike’s Information Criterion (AIC) of the whole\textsuperscript{26}. We assessed the explanatory significance of the AIC-minimizing model components by comparison to the fit obtained when using no exponential smoothing for the relevant subcomponent (trend or seasonality) when holding all other methods identical. We also calculate the seasonal and non-seasonal variants of the Mann-Kendall test for trend, whose tau statistic provides a guideline as to the direction of trend where applicable\textsuperscript{27}. For consistency between groups, we used the measure of patient count relative to all patients with any diagnosis record in the filtered dataset.

**Results**

We identified 11 statistically significant coding trends that were deemed to be of practical interest. These trends are summarized in Table 2. Each of the novel ICD-10 codes identified in the reported trends can be categorized in one of two ways. First, novel codes may relate specifically to body site laterality – left knee or right shoulder, for example. Alternatively, novel codes may represent an increase in granularity over similar ICD-9 codes.

Table 2. Summary of coding trends, by novel ICD-10 code and corresponding header. Code type describes our categorization of what makes this code distinct from related ICD-9 codes. Percent change is the absolute change in the relative patient share of the novel code(s) over the entire study period.

<table>
<thead>
<tr>
<th>ICD-10 Code</th>
<th>Description</th>
<th>Header</th>
<th>Description</th>
<th>Novel Code Type</th>
<th>Δ %</th>
</tr>
</thead>
<tbody>
<tr>
<td>E11.65</td>
<td>Type 2 DM with hyperglycemia</td>
<td>E11.6</td>
<td>Type 2 DM with complication</td>
<td>Granular</td>
<td>-10.8%</td>
</tr>
<tr>
<td>Z79.891</td>
<td>Long term use of opiate analgesic</td>
<td>Z79.81</td>
<td>Long-term drug therapy</td>
<td>Granular</td>
<td>-2.8%</td>
</tr>
<tr>
<td>I25.5</td>
<td>Ischemic cardiomyopathy</td>
<td>I25</td>
<td>Chronic heart disease</td>
<td>Granular</td>
<td>+5.8%</td>
</tr>
<tr>
<td>Z12.31</td>
<td>Mammogram for breast cancer</td>
<td>Z12.3</td>
<td>Breast cancer screening</td>
<td>Granular</td>
<td>+9.9%</td>
</tr>
<tr>
<td>R09.81</td>
<td>Nasal congestion</td>
<td>R09.8</td>
<td>Respiratory symptoms</td>
<td>Granular</td>
<td>+7.2%</td>
</tr>
<tr>
<td>M25.51X</td>
<td>Pain in right/left shoulder</td>
<td>M25.51</td>
<td>Pain in shoulder</td>
<td>Laterality</td>
<td>+8.4%</td>
</tr>
<tr>
<td>M25.55X</td>
<td>Pain in right/left hip</td>
<td>M25.55</td>
<td>Pain in hip</td>
<td>Laterality</td>
<td>+8.9%</td>
</tr>
<tr>
<td>M25.56X</td>
<td>Pain in right/left knee</td>
<td>M25.56</td>
<td>Pain in knee</td>
<td>Laterality</td>
<td>+7.0%</td>
</tr>
<tr>
<td>M79.60X</td>
<td>Pain in right/left leg</td>
<td>M79.60</td>
<td>Pain in limb</td>
<td>Laterality</td>
<td>+9.7%</td>
</tr>
<tr>
<td>M17.1X</td>
<td>Osteoarthritis of right/left knee</td>
<td>M17.1</td>
<td>Osteoarthritis</td>
<td>Laterality</td>
<td>-3.7%</td>
</tr>
<tr>
<td>D50.1</td>
<td>Sideropenic dysphagia</td>
<td>D50</td>
<td>Iron deficiency anemia</td>
<td>Other/unspec.</td>
<td>-</td>
</tr>
</tbody>
</table>
Although this work is primarily focused on changes in the usage of novel ICD-10 codes, our modelling approach also detected the tradeoff between non-novel codes closely related to codes of interest. In Table 2, we indicated this coding pattern as ‘other/unspec.’ to indicate that the calculated statistics typically reflect differential usage of the other and unspecified code types, rather than changes in usage of the new, more specific codes. For example, within the header grouping D50, ‘Iron deficiency anemia,’ the new more granular code of interest is D50.1, ‘Sideropenic dysphagia.’ Despite the relative consistency of D50.1 coding over the study period, there was a significant alteration in the comparison baseline of D50.8 and D50.9 (other anemia and unspecified anemia, respectively). We present this trend because decreased usage of ‘unspecified’ codes in favor of ‘other’ codes may indicate an area where the clinical content coverage of ICD-10 is lacking.

One of the most common types of coding pattern identified through our analysis was the increased use of laterality codes over time. For example, Figure 1 provides a clear example of reduced usage of unspecified laterality over time for the code header M25.51, pain in shoulder. This increased uptake of laterality codes is most pronounced for pain related codes; we observed increases in laterality-specific coding within the code groups describing pain in the shoulder, hip, knee, and limbs. We did also note a decrease in usage for M17, osteoarthritis of the knee.

![Figure 1. Use of laterality codes for M25.51 header (pain in shoulder)](image)

Generally, the novel codes for laterality behaved similarly to one another over the study period. There was more diversity among the trends identified for novel codes that were more specific than their ICD-9 counterparts. Of the five identified trends involving more granular ICD-10 codes, three were for increased usage of the novel codes and two were decreases in usage.

We also identified certain trends that had a dramatic seasonality to the relative usage of novel ICD-10 codes. An example of a highly-seasonal trend is provided in Figure 2. For the most part, our analysis is not susceptible to the underlying seasonal trends of disease and healthcare utilization because we are primarily interested in the relationships among codes that share similar underlying temporal attributes. However, in cases where there are clearly different month-to-month behaviors for the different codes under a single header, our analysis will include artifacts from that
seasonality. After correcting for this seasonal behavior, there remains a more subtle trend which we still present as a valid shift in coding patterns.

![Figure 2. Seasonal coding patterns for R09.8 (respiratory symptoms)](image)

**Discussion**

We expected to see an increase in novel ICD-10 code usage over time as a result of increased familiarization. The increased uptake of laterality information is an excellent example of such a positive learning curve. Initially, coders may have been unaware of the increased documentation possibilities or were using ICD-9 crosswalks which would map to unspecified laterality codes. The adoption of these laterality codes represents a positive development: usage of these new ICD-10 codes is likely to improve care and reduce fraud.

The discrepant laterality trend, M17 - osteoarthritis, was the single reported decline in usage of the laterality codes. We suspect that this trend is due to a data artifact that does not represent actual coder behavior. Over a 3-month period in early 2017, there was a substantial jump in the raw number of unspecified laterality variant of arthritis codes. This coincided with an overall increase of arthritis codes in our analysis and is likely due to a newly-introduced data contributor to the overall dataset. Contributors are deliberately concealed in this deidentified dataset, and the only proxy for determining such effects is through analysis of changes in geographical distributions of patients in the data. Although such increases in raw numbers was rare, further work will be required to fully understand the impact of such distributional shifts.

Apart from the usage of laterality codes, the perceived trends for new ICD-10 codes were of mixed magnitude and direction. Some of the reported trends were straightforward increases in adoption of more specific codes: Z12 added details about mammograms, I25 added a code for cardiomyopathy, and R09 included increased symptom specificity. However, some of the trends moved in the opposite direction: the header groupings for drug therapy and complicated type 2 diabetes (T2DM) saw shifts in usage away from novel codes and towards more unspecified or other codes. For example, in the T2DM header E11.6, the usage of the newer more granular code E11.65 (hyperglycemia) decreased and the generic, unspecified code E11.69 became more popular. See Figure 3 for the patient shares of these codes over time. There are a number of possible explanations for this type of trends: they may be explained by an evolution of diagnostic practice that ICD-10 fails to cover.
In addition to the temporal anomaly observed in the distribution of the osteoarthritis codes mentioned previously, there were some other data fluctuations on short timescales. For example, the coding for breast cancer screening experienced a noticeable discontinuity in August 2017, where the use of the novel mammogram code Z12.31 increased about ten percent among patients receiving any code for such a screening. There are several potential explanations for such a shift; to our knowledge no new screening guidelines were issued around that period, but hospital systems might have been adopting 2016 guidelines at that time. Alternatively, coders may have been responding to reimbursement pressures around screening codes. No other trend experienced a short-term shift as large as that of the breast cancer screening header, and the others may be attributable to random noise.

**Conclusion**

Understanding the uptake of ICD-10 codes is critical to the accuracy of ICD-based analyses. Our work has identified several broad categories of coding patterns that should be corrected for or excluded from retrospective studies. The introduction of laterality codes in ICD-10 represents an opportunity for more detailed study of the specific body sites of given conditions. However, researchers need to be aware of the learning curve effects present in these codes after the ICD-10 transition. Similarly, the increased specificity for screening encounters available in ICD-10 enables enhanced study of utilization and health system effects but may be affected by the trends identified here. Finally, the cases of decreased uptake of novel codes may indicate areas where revisions or additions to ICD-10 are necessary and should be tracked in future ICD-10 implementation studies.

There are several opportunities for more in-depth analysis of key coding trends revealed by this work. Throughout our trend modelling, we restricted code groupings to the header codes provided by CMS. Although sufficient for our purposes, there are higher level groupings that might reveal broader shifts in coding practice. Looking at coding practices across higher-level ICD-10 groupings or other clinically significant clusters would present highly relevant patterns to coding and informatics professionals. Additionally, novel features of ICD-10 can be spread across many different code groups. Analyzing the usage of cause of injury details or overlapping site codes would provide useful learning curve information. Another factor to be considered in future work would be the “grace period” that CMS applied only to the first fiscal year (2016) of ICD-10 implementation, which allowed for a degree of imprecision in Medicare fee-for-service claims; the freedom from audit that it promised in that year could help to explain some of the variation in coding patterns observed between that year and fiscal years 2017 and 2018. Finally, this work focuses
mainly on novel ICD-10 codes but, as we move further away from the ICD-10 transition, that novelty may prove less relevant to coding practice and our analysis should be expanded to other properties of ICD-10 codes and the relationships among them.

For future coding system updates, including both structural and incremental changes, this type of coder uptake analysis is vital to understanding the impact of various incentives on reported outcomes. This knowledge of coding practices is important not just to coders and researchers; it suggests important features of the inner workings of our health care systems.

References
6. Fox B, Sheehan J. Openness and Exactness - Mitigating Fraud Vulnerabilities in the Age of EHRs and ICD-10 [Internet]. HIMSS; 2012.
810
25. IBM Explorys Cohort Discovery, IBM Explorys Therapeutic Datasets, and IBM Explorys Virtual Workbench provide life sciences insights into real-world care delivery. 2016 Sep 13. IBM United States Software Announcement 216-401.


28. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The Content Coverage of Clinical Classifications. JAMIA. 1996 May 1;3(3):224–33.


Parent Perspectives on Pediatric Inpatient OpenNotes

Catherine Arnott Smith, PhD1; Ryan J. Coller, MD, MPH2; Shannon M. Dean, MD2,3; Daniel Sklansky, MD2; Peter L.T. Hoonnaker, PhD4; Windy Smith, MSN, RN, CPN3; Anne S. Thurber, MS2; Brad D. Ehlenfeldt, BBA3; Michelle M. Kelly, MD2,4

1The Information School, University of Wisconsin-Madison, Madison, WI; 2Department of Pediatrics, University of Wisconsin School of Medicine and Public Health, Madison, WI; 3UW Health, Madison, WI; 4Center for Quality and Productivity Improvement, University of Wisconsin-Madison, Madison, WI

Abstract

Introduction: 169 U.S. health systems now engage in OpenNotes: a movement to share clinical notes with patients. Few studies have focused on releasing notes during hospitalization, pediatrics, or parents/caregiver perspectives.

Methods: A focus group was conducted with eight parents with experience caring for a hospitalized child at a Midwest children’s hospital. In the 2-hour session, parents were asked about their perspectives of the idea of sharing inpatient doctors’ daily notes with parents during their child’s hospitalization. Qualitative analysis was conducted to elicit themes related to the potential benefits and challenges of sharing inpatient notes. Results: The most mentioned benefits included notes providing information as a reference for improved family education/understanding, communication/continuity, and advocacy/empowerment. Challenges were primarily related to note content, impaired communication and negative impact on families. Conclusion: Participants identified multiple potential benefits of and challenges to sharing notes with parents during their child’s hospitalization but also acknowledged the impact on healthcare professionals who work alongside them.

Introduction

The OpenNotes movement, launched in 2010, advocates “fundamental change in the way visit notes are managed”, by making visit notes available to patients.1 The objective of this overarching study was to identify stakeholder perspectives on the idea of sharing inpatient doctors’ daily notes with parents during their child’s hospitalization using an inpatient portal, MyChart Bedside (Epic Systems). We conducted five focus groups composed of different stakeholders at a quaternary children’s hospital in the Midwest: (1) parents (Patient and Family Advisory Council members with experience caring for a hospitalized child); (2) nurses (pediatric bedside nurses working on general medical/surgical units); (3) pediatric intern and resident physicians; (4) attending physicians; and (5) representatives from the departments of hospital administration, information services, risk management and patient relations. Findings from this study will be used to inform design requirements for optimally implementing and evaluating OpenNotes for hospitalized patients and caregivers using an inpatient portal. This paper reports on some of the findings from the Parent focus group.

Background

In 1973, the psychiatrists Shenkin and Warner published an editorial in the New England Journal of Medicine proposing that patients should be given complete copies of their medical record and routine issuing of new copies as information changed. This ignited a firestorm of dialogue among healthcare professionals.2 It has been called the “pioneering” work in the field of patient access.3 Alan Westin, a pioneering advocate for consumer data privacy, ascribed the patient access initiative to a developing “consumer” theory of health care.4 This theoretical lens saw patient access as beneficial for patients and the larger healthcare system alike by improving content, patient education, and patient-physician communication. Patient readers could correct erroneous information found in the record, thus ultimately enhancing the quality of care.4 Record access could also be a part of continual patient self-education, enabling smarter health consumerism and better client-physician relationships.5,6 Patients could be “more active partners” if they were allowed more information about their treatment.2,4 Further, medical record access could result in a “realignment of power.”7 Objections to such access by physicians fell into dominant themes, which Westin called...
the “doctors’ judgment model.”4 Opponents located their objections in the nature of the clinical documents themselves. In terms of content, it was argued that reading medical records could result in actual harm to the patient through causing confusion;8 of education, that the amount required to teach a patient to understand what they read placed a burden on the physician;4 and that communication between physician and patient would be negatively affected by medical record access.9

The question of how much, if any, of their medical records patients and/or their caregivers should be able to see has also been a longstanding debate in healthcare. It is no accident that the first published study of the effects of access, which appeared the same year as Shenkin & Warner’s editorial, was done in a context of Problem-Oriented Medical Record development.10 Health informatics technologies have the potential to greatly affect the balance of power between physician and patient simply by providing the patient with more information, which requires attention to knowledge representation for new readers as well as re-education of the writers.

In 2010, a movement called OpenNotes began as a collaborative experiment between hospitals in Boston, Pennsylvania and Seattle. Twenty thousand patients were given the opportunity to read their clinical notes – textual summaries of their doctor visits – using their patient portals. In two of the study sites – Pennsylvania and Boston – 82% of patients opened the notes, while in Seattle only 47% did. The response from patients was overwhelmingly positive. A very high proportion (77-87%) agreed that the notes helped them feel “more in control of their care”. By self-report, 60-78% reported increased adherence to their medications.11 As a result of this early success, as of March 10, 2019, 169 health systems in the U.S. “and a couple in Canada” are now engaging in OpenNotes.12

Themes consistent with the consumer movements of the 1970s continue to arise in the 21st century research literature of OpenNotes. Benefits include enhanced communication (text supplements to verbal instructions about care plans and appointments),13 improved patient knowledge and potentially improved self-care (understanding of diagnosis and care plans; patient activation; error correction),13,14 and greater patient participation (medication adherence, monitoring and follow-up with physician).13,14 Challenges include concerns about security,15 stress induced by reading notes,13 information overload,15 confusion induced by terminology,13,16,17 unintentionally offensive language,13 and technical issues.13

The research literature explicitly referencing the OpenNotes initiative is growing with the movement, but a PubMed search reveals only 20 published studies in which patients were involved and patient feedback sought. Of these, 15 were online surveys or analysis of secondary data from surveys, for example.18 The remaining five studies used qualitative methods, such as semi-structured interviews or focus groups. Surveys have been done of 13 patients in primary care;19 28 patients living with mental health conditions;20, 21 20 oncology patients;22 and 14 patients seen at a university health service.23 One of the numerous My HealtheVet studies examined the patient experience using focus groups of patients inside and outside the VA.24 Of note, only two studies have involved caregivers of patients. Bell ET al.14 reference unpublished studies of focus groups conducted during the early years of OpenNotes that included caregivers. Wolff et al. comment: “Family members and friends are frequently involved in the care of patients who are more vulnerable and may facilitate access and use of a patient portal. However, the longitudinal experiences of patients and families who share formal access to a patient’s portal account has, to our knowledge, not been examined.”

Nowhere is the involvement of caregivers clearer than in the domain of pediatrics. OpenNotes reported that 76 healthcare systems were sharing pediatric notes with patients, parents, or both.26 However, Sarabu et al. call attention to the complexities posed by pediatric care, “unique because additional caregivers ... are almost always involved and generally have legal decision-making capacity for most medical decisions that involve the patient.”27 The objective of this study was to identify parent perspectives of the idea of sharing doctors’ notes with parents during their child hospitalization. Qualitative research is required to understand the diverse and nuanced perspectives of stakeholder participants. This is especially important when considering changes within complex healthcare systems, like the implementation of OpenNotes in the pediatric hospital setting. This paper reports the findings from a focus group composed of important pediatric participants: parents.

Methods

Setting and participants. In the overarching study, focus groups were conducted with each of the 5 stakeholder groups at an academic children’s hospital in the Midwest during October and November 2018. The study reports on findings from the parent focus group which met once. Eligible parent participants were members of the hospital’s Patient and Family Advisory Council and were invited by email. Respondents were provided an information sheet describing the
study, risks and benefits, and contact information, and informed consent was obtained. Participants were not reimbursed for their participation. The institution’s Institutional Review Board approved this study.

**Study procedures.** The parent focus group was conducted in a private hospital conference room. The facilitator was a research specialist trained in qualitative methods (AT). A semi-structured facilitator guide was developed and refined after piloting with stakeholder representatives. A hybrid approach was used integrating both open-ended and targeted questions. Focus group sessions started with a description and definition of inpatient doctor’s daily progress notes and followed with open-ended probes to explore participant perceptions of the idea of and benefits of and challenges to sharing notes with parents during hospitalization through the inpatient portal. At the end of the session, the facilitator provided a summary, offered an opportunity for revisions or clarifications and administered a short demographic survey to participants. The facilitator guide can be found at: www.hipxchange.org/InpatientPortal.

**Data analysis.** Audio-recordings were transcribed by a professional service. A ‘cleaning’ process was then performed to ensure accuracy, fill in text that the transcriptionist could not discern, and delete any identifying information inadvertently collected. To ensure anonymity, recordings were destroyed after transcription. Transcribed data were transferred to Dedoose© (https://www.dedoose.com), software used to facilitate the organization and coding of qualitative data.

Three researchers trained in qualitative methods participated in thematic content analysis of the focus group transcript (AT, CAS, MK). Data was organized into themes related to benefits and challenges and the analytical approach was based on this framework. Two researchers (AT, CAS) first independently reviewed the transcript, identifying themes and concepts emerging from the data related to benefits and challenges to share inpatient doctors’ notes with parents during hospitalization. With a third researcher (MK), they developed and iteratively refined a codebook of identified themes, their definitions and exemplar quotes. Two researchers (AT, CAS) then independently coded the transcript. During this process, all three researchers continued to meet together to review the coding and reach consensus over any discrepancies, always referring back to the transcript.

**Findings**

The Parent focus group discussed here had 8 participants and lasted 2 hours. All but 1 participant was female, and ranged in age from 35 to over 55 years old. Half had a college degree, 75% (N=6) were White and 25% (N=2) Asian. All reported that they spent time on the Internet, using a computer, a tablet, or smart phone. The majority (N=6, 75%) more than 15 hours a week, 12% (N=1) between 6 and 10 hours a week, and 12% (N=1) between 1 and 5 hours a week.

The most mentioned themes related to potential benefits focused on notes as a means of information provision: serving as reference for improved family education/understanding, communication/continuity and advocacy/empowerment. These themes and examples are shown in Table 1.

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Parent quote</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>REFERENCE FOR IMPROVING FAMILY EDUCATION/UNDERSTANDING</strong></td>
<td></td>
</tr>
<tr>
<td>Provide a recap of information, reinforcement, supplement or reference for family</td>
<td>So I really hope that this kind of comes to being, because it will give, especially patients that have long-term care, more access. And having that at your fingertips is ... so much easier when you have to kind of remember down the line something for the school, or something for a social worker, or something like that that you could can quick go back and look at ... where was he on this scale when he did his neuro test, you know?</td>
</tr>
<tr>
<td>Improve family understanding or knowledge</td>
<td>I've had the good fortune to see some of the doctor's notes, and like you said earlier, I think it allows you to learn a little bit more about what's going on and maybe do a little checks and balances along the way. Did so-and-so understand, or did I understand what was being said?</td>
</tr>
</tbody>
</table>
Table 1, cont’d.

<table>
<thead>
<tr>
<th><strong>ENHANCE COMMUNICATION/CONTINUITY</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Improve parent-provider and/or team</strong></td>
</tr>
<tr>
<td>communication</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

| **Ensure continuity of information/care** |
| between providers                       |
|                                     | *Without a tool or an avenue like that, there are pieces of the puzzle that sometimes get missed. And one of the times we really noticed it was on weekends, when staff maybe had changed over or there weren't so many of the other ancillary staff, but it was sometimes really hard to get information. And nobody was even sure where to go to get that information, so it seems like it's a really good place to centralize a lot of that.* |

<table>
<thead>
<tr>
<th><strong>PROMOTE ADVOCACY/EMPOWERMENT</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Improve empowerment, engagement, control for parent or child</strong></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

| **Parent advocacy for child** |
|                                     | *I also think that there are parents that might not be so inclined to be involved. But if the tool is present, it might encourage them to do something like that that they might otherwise tend to shy away and wait for somebody else to tell them everything.* |

| **Family identify errors/inaccuracies** |
|                                     | *I'm a parent of an 11-year-old. At one point, he was in the hospital pretty continuously. But we've been lucky, and he's been mostly outpatient but still come in for procedures and surgeries and stuff at least twice a year. So we have a lot of experience with a lot of the different specialties, and I think when he does get admitted, we do like to be able to have the continuity of care. Sometimes, you know, there might be critical pieces of information that may not have been stressed enough or could be missed in that period of rounds, and so it gives you the opportunity to say, hey, this other topic you know, that was really important to me. And only with a second set of eyes would you be able to capture that information, so I think it's really important.* |

Challenges identified by parent focus group participants constituted 37 distinct themes, with no one theme comprising more than 11% of all comments. Top themes were related to note content and the potential negative effect of sharing notes on communication and family interactions. These themes and examples are illustrated in Table 2.
<table>
<thead>
<tr>
<th>Challenges</th>
<th>Parent quote</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NOTE CONTENT</strong></td>
<td></td>
</tr>
<tr>
<td>Medical jargon</td>
<td>It's not just the language, there just needs to be somebody that people can ask questions. Because it isn't even just medical words. It's like Q4HPRN, you know, every four hours, PRN. Does everybody know what that means? So that's not a word. That's just medical abbreviations. Maybe everybody in this room has had enough experience that they know what that is, but I would venture to guess that there are a lot of people that that would just be a jargon for.</td>
</tr>
<tr>
<td>Discussion of sensitive diagnoses (e.g., eating disorder, mental health, abuse) or private topics (e.g., sexual health, social issues, family dynamics) in notes</td>
<td>The other challenge I was wondering about is if there are certain things a patient wants to keep private… I think this came up several times about use of birth control. Personally, I don't think it's my personal opinion as a parent that if my child is on birth control, I need to know. But I would respect the law, whatever it is. So now does this open the door that like if it goes in the notes, then now someone has access to that?</td>
</tr>
<tr>
<td><strong>IMPAIRED COMMUNICATION</strong></td>
<td></td>
</tr>
<tr>
<td>Release of information (e.g., abnormal tests, diagnosis) before face-to-face communication or without translation by provider</td>
<td>I suspect that notes will also reflect comments between providers that parents are not yet aware of. I'm thinking it may not just be what happened at rounds that we know already, but after rounds, a consultation was had or some discussion was had, and that's going to show up in the notes, and we have not yet become privy to that, so the whole timing issue … brings that issue up again.</td>
</tr>
<tr>
<td>Physicians inhibited from communication with each other</td>
<td>If I really trusted my kids' doctors, I want them to go in that room without me and to play with their creative, sciency brains and go, “It could be this, it could be that.” I don't want to hear that. … I want them to figure it out and then say, hey, we're strategizing this. … I would hate for them to somehow stifle that part of their collegiality, because, oh, mom is going to read it.</td>
</tr>
<tr>
<td><strong>IMPACT ON PATIENT OR FAMILY</strong></td>
<td></td>
</tr>
<tr>
<td>Negative parent emotion induced by information in note (e.g., frustration, confusion)</td>
<td>We got called into a conference room with, what, like four doctors, four nurses, a couple social workers. I mean, we had the whole squad there, you know, and they basically told us we had a 5% chance of survival, and that we could take our daughter home and put her on hospice. If I would have read that in a doctor's note, no matter good I am at absorbing this stuff, I think would have probably freaked out then too, because it was not a good day.</td>
</tr>
<tr>
<td>Parents/families are different</td>
<td>Other parents are going to handle [access to notes] differently. We might be, you know, more professional, or we might be upset but then like keep that poker face on. But some families would probably create quite a bit of static over certain things and stifle that.</td>
</tr>
</tbody>
</table>

Some comments did not fall neatly into the Benefit or Challenge themes. On numerous occasions parent participants seemed concerned about the impact of inpatient OpenNotes on other stakeholders:

“I'd be shocked if the doctors really wanted it.”

The central theme of these comments was concern for healthcare professionals’ time – time that parents wanted to be spent on caring for their children and communicating with other healthcare professionals:

“I don't want to make their jobs more difficult as doctors, and I don't want to burden the nurses and other medical staff with all the questions that this could bring up. I mean, I'm very sensitive to the nurses' time.”

“I don't ever want to interfere with them being able to do their job.”
Discussion

This is the first study to evaluate parent perspectives of the potential for OpenNotes in the pediatric inpatient setting. There is a striking resemblance between these parents’ perceptions of sharing notes and the potential of such sharing enunciated almost 50 years ago at the dawn of the patient access initiative, a focus on content, education, and communication. These themes are echoed in recent studies evaluating patient perspectives of ambulatory OpenNotes and a pilot study assessing inpatient note sharing with 10 hospitalized adults.

These results also highlight multiple potential benefits and challenges that may be unique to sharing doctors’ notes in pediatrics and in the hospital setting. Parents mention characteristics of the child or family that may make note sharing more challenging, such as with hospitalized children approaching adolescence where it is vital to ensure that their information is kept private. Parents’ worry and anxiety may be heightened in the fast-paced, high acuity inpatient setting. Participants suggest this anxiety may be worsened with access to notes, particularly if information is shared prior to face-to-face communication with physicians. On the other hand, note sharing may facilitate parent communication with their child’s inpatient physician. For example, one parent suggests that the information found in notes may support parent engagement in communication with physicians during daily bedside rounds. Another mentions that access to notes may help parents act as “a second set of eyes” to ensure information accuracy and continuity specifically during inpatient care transitions, such as during staff change of shift. These results suggest that there may be differences in the implementation, use and impact of OpenNotes in the inpatient setting, which will be important areas for future research.

The opinions of parent participants in this focus group also reflect parents’ awareness of the potential impact of note sharing on the rest of the healthcare team. They identified potential benefits for parents during their child’s hospitalization, but also acknowledged the potential impact on healthcare professionals. In fact, one of the most frequently expressed ideas during the focus group session by these parents included comments about the potential effects of note sharing for providers. One benefit, for example, was “ensuring continuity of care/information between providers,” and a challenge was the possibility that “physicians [would be] inhibited from communication with each other.”

These results will inform efforts to design, implement and evaluate OpenNotes in pediatric inpatient settings. Potential benefits identified by parent focus group members here will be used to define meaningful outcomes for families. Example outcomes may include improved parent-physician shared understanding upon discharge and family identification of medical errors. Challenges will also be considered proactively and efforts made to mitigate potential negative consequences prior to the wide-scale implementation of inpatient OpenNotes. Parents in this focus group highlight the rapidly changing nature of inpatient care and risk of discordance between the information communicated to parents and updated and shared in notes. Organizations may need to set expectations for parents regarding the fluidity of care plans and who to contact if and when questions arise. Strategies such as these are necessary for hospitals to jointly support the work of patients, caregivers and their healthcare teams.

This study has limitations. Results reflect the perceptions of a small number of parent participants from an academic children’s hospital who may represent a more engaged population of parents. Consistent with the intent of exploratory research, the findings of this qualitative study were not intended to be generalizable, but instead lay the groundwork for future research evaluating the growing use of inpatient portals and expansion of the OpenNotes movement into the inpatient setting.

Conclusion This is the first study assessing the potential to share inpatient doctors’ daily notes with parents during their child’s hospitalization. Parents anticipated multiple potential benefits and challenges of inpatient note sharing, including the impact of note sharing on healthcare providers who work alongside them. Some themes from this parent focus group are consistent with the benefits and challenges described in the literature on outpatient notes sharing, but others highlight concepts that may be unique to pediatric and inpatient note sharing. Findings from this study will inform the design and implementation of BedsideNotes, an intervention and implementation strategy to share inpatient doctors’ daily notes with families to support their engagement in hospital care.

ACKNOWLEDGMENTS

This project was supported by the NIH CTSA at UW-Madison grant 1UL1TR002373, and the University of Wisconsin-Madison School of Medicine and Public Health’s Wisconsin Partnership Program, WPP-ICTR grant # 3086.
References


about safety-related knowledge, behaviors, and attitudes after reading visit notes. J Patient Saf. 2018 Apr 27. doi:10.1097/PTS.0000000000000494


Comparing Deep Learning Models for Multi-cell Classification in Liquid-based Cervical Cytology Images

Sudhir Sornapudi, Ph.D. student¹, Gregory T. Brown, M.D., Ph.D.², Zhiyun Xue, Ph.D.², Rodney Long, M.A.², Lisa Allen, B.S., CT³, Sameer Antani, Ph.D.²
¹Missouri University of Science and Technology, Rolla, MO, USA; ²Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, USA; ³Diagnostic Systems Women’s Health and Cancer, Becton Dickinson and Company, Durham, NC, USA

Abstract

Liquid-based cytology (LBC) is a reliable automated technique for the screening of Papanicolaou (Pap) smear data. It is an effective technique for collecting a majority of the cervical cells and aiding cytopathologists in locating abnormal cells. Many methods published in the research literature rely on accurate cell segmentation as a prior, which remains challenging due to a variety of factors, e.g., stain consistency, presence of clustered cells, etc. We propose a method for automatic classification of cervical slide images through generation of labeled cervical patch data and extracting deep hierarchical features by fine-tuning convolution neural networks, as well as a novel graph-based cell detection approach for cellular level evaluation. The results show that the proposed pipeline can classify images of both single cell and overlapping cells. The VGG-19 model is found to be the best at classifying the cervical cytology patch data with 95% accuracy under precision-recall curve.

Introduction

Cervical cancer is the second most common cancer in women living in under-developed regions. Nearly 570,000 new cases were recorded in 2018 and about 311,000 women died from cervical cancer worldwide¹. In the United States (2019) it is estimated that about 13,170 cases will be diagnosed for invasive cervical cancer and about 4,250 women will die from cervical cancer². Fortunately, cervical cancer can be treated successfully if detected at early stage. LBC³ for Pap (Papanicolaou) test is the gold standard for cervical cancer screening and has significantly contributed to reducing mortality. However, manual examination for detecting abnormal cells in a cervical cytology slide is a tedious process even for an expert cytologist. Expedient secondary reviews are conducted in areas marked by the pathologist or cytotechnologist with an ink marker by hand. There is a need for automated and computer-assisted technique for fast and efficient screening.

Though, generally abnormal cells have a relatively higher nuclei to cytoplasm ratio within a cell body⁴, it is very time consuming and requires significant training and expertise to manually locate these abnormal cells under a microscope. While there are some automated approaches such as Becton-Dickinson’s FocalPoint⁴ and Hologic’s ThinPrep⁵, both manual and automated cytology are challenged by the high variability in cell size, shape and color, and complex morphology due to overapped or crowded cells. In recent years, computer-assisted automatic approaches have shown promising results in cell classification⁶. The current deep learning era has vastly improved the performance and the classification accuracy in various biomedical applications. While some methods avoid pre-segmentation step, much work in the literature shows considerable research in the direction of cell segmentation of cervical cytology images and single cell classification. Segmentation using superpixel-wise convolutional neural network with dynamic shape modelling was employed by Tareef et al.⁷ Watershed⁸ and contour-seed pairs learning-based framework⁹ were some of the successful approaches in segmenting overlapping cells. However, accurate segmentation of cervical cells is impeded by overlapped and clustered cells. Cell classification is the next step after segmentation. Pixel-level classification was employed by extracting traditional features and training an SVM classifier¹⁰. Similarly, block-level classification with SVM classifier was studied¹¹. DeepPap¹² proposes CNN based cell classification on cell image patches, which are handpicked, cropped and centered at nuclei. All these approaches consider images containing only single cell or slightly overlapped cells, which are far from the real-world conditions. Traditional and machine learning

---

methods were reviewed and discussed\textsuperscript{12} for automated cervical screening from the pap-smear images. Our work lays foundation based on the stages involved in automated pap-smear analysis as mentioned in the literature review\textsuperscript{12}. The stages for analysis typically comprise of image acquisition, preprocessing, identifying abnormal regions, feature extraction and finally classification.

**Figure 1. Framework of the proposed approach**

The main objective of our study is to classify and locate the abnormal cervical cells. Understanding the challenges and limitations of the previous approaches, we aim to extract cell image patches from microscopic images under realistic conditions where cells may overlap; we evaluate different CNN models toward classifying the image patches.

**Materials and Methods**

The objective of the study is to generate clean labeled image data from the multi-level microscopic whole slide images to classify the cell image patches into normal and abnormal. First, raw cytology slide data (both clean and inked data) is preprocessed to obtain regions containing abnormal cells. The preprocessing comprises alignment of inked image data with respect to the clean image data through feature-based image registration. Lower resolution image levels (level 7) were used to speed up the process. Next, these lower resolution image levels were analyzed to obtain the ROI bounding box coordinates. This is accomplished by simple image processing techniques like image subtraction, thresholding, applying morphological operations and skeletonizing ink identified masks to detect and refine bounding boxes. Then, annotated image data is generated from the high resolution (level 1) images through accurately drawn abnormal cell boundaries on the ROIs. Finally, the labeled data was combined with Herlev Pap smear dataset\textsuperscript{13} to train and test various CNN architectures and later evaluate the results to choose the best CNN model that classifies cytology data. A novel graph-based cell detection model was also proposed to identify the cell boundaries even under overlapping cases in the high-resolution image regions. The approach includes over segmenting the image with superpixels and connecting the centroids to create a graph and then identifying sub-graphs (cell regions) with respect to nuclei nodes obtained by applying appropriate graph cuts in the graph. The flow of this proposed approach is illustrated in Figure 1.

**Datasets**

The study uses two datasets. The first set comprises 25 cervical liquid-based cytology slides provided by Becton-Dickinson (BD) Corporation using their Sure Path technique\textsuperscript{14}. In the case of abnormal slide data, there are a pair of slide images from each patient. One image contains only clean slide and the other image contains blue ink marks annotated by an expert cytotechnologist to indicate the regions containing abnormal cells (later, abnormal cells from
these regions are identified by an expert pathologist). The ink marks can be easily cleaned with an alcohol swab. The glass slides were scanned both with the marks, and after their removal, using a Hamamatsu NanoZoomer 2.0-HT whole-slide scanner, producing digitized slides in a pyramid tiled format with the file extension ndpi. The NDPI files resulting from the scanning are large in size (100’s of megabytes of size data). The second dataset is the publicly available Herlev Pap Smear dataset (http://mde-lab.aegean.gr/downloads) where specimens are prepared via conventional Pap smear. The dataset contains 917 single cervical cell images. Table 1 shows the characteristics of the data used for the study.

Table 1. Characteristics of the datasets used for the study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image type</th>
<th>Pixel Size (in μm)</th>
<th>#Normal</th>
<th>#Abnormal</th>
<th>File type</th>
<th>Total Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD Corp. Data</td>
<td>Whole slide image</td>
<td>0.228x0.228</td>
<td>6</td>
<td>19</td>
<td>NDPI</td>
<td>25</td>
</tr>
<tr>
<td>Herlev Data</td>
<td>Single cervical cell</td>
<td>0.201x0.201</td>
<td>242</td>
<td>675</td>
<td>BMP</td>
<td>917</td>
</tr>
</tbody>
</table>

**ROI Detection**

While the entire cytology slide sample contains thousands of cells, just a few abnormal cells are sufficient indicators of abnormal screening. These abnormal cells may be surrounded by a large number of normal cells. Relative morphological, nucleic, or cytoplasmic appearance differences between cells helps pathologists to identify them. These regions are marked in ink on the glass slide. We use these markings on scanned slide image as region of interest (ROI) indicators. It is important to note that not all regions containing such abnormal cells were marked nor were individual abnormal cells identified within inked regions.

The NDPI format is organized as a pyramid structure with multiple levels of down-sampled subimages. A low-resolution image of level 7 is found suitable for easy preprocessing for detecting ROI bounding box coordinates. We correct any uniplanar misalignment problems between the blue ink-marked and ink-removed slides using image registration through feature based image alignment. The clean slide image is considered as a reference image, and stable ORB (Oriented FAST and Rotated BRIEF) feature points along with descriptors are obtained. These key point features are mapped with the features in the inked slide image (target) as shown in Figure 2. Homography is calculated based on the mapping information using Random Sample Consensus (RANSAC) estimation technique. The transformation is finally applied on the target inked slide image to map it to the reference clean image.

The preprocessing step also includes subtraction of blue (ROI ink) color space from red color space, which makes it easy to create a threshold ROI binary mask in lower resolution image. Skeletonizing and refining boundaries generated accurate ROIs. The coordinates of these ROIs were normalized and recorded. These ROIs are then cropped out from the high-resolution clean slide image (level 1) using the normalized bounding box coordinates data. Figure 3 shows the resultant intermediate output images.

**Figure 2.** Matching keypoints in a low resolution image to align annotated slide image to the reference clean slide image

**Figure 3.** The resultant intermediate output images.
We propose two methods for cell data generation from the extracted ROIs: Graph-based cell detection and patch-based data generation.

**Graph-Based Cell Detection:**
We chose this method to generate one cervical cell image (Herlev-like dataset). Initially the ROI image is over-segmented by generating superpixels through QuickShift approach\(^\text{\scriptsize 17}\). The pixel intensities are averaged over each superpixel region. The resultant image is converted into a graph with centroid of superpixels as nodes, line connecting the adjacent nodes as edges and absolute difference of L2 normalized color intensities at the respective adjacent nodes as edge weight. Graph cut is used at an empirically determined threshold of 59. The resultant image is a binary mask for the nuclei present in the ROI image which form the graph nodes. The novelty of the proposed approach lies in generation of subgraphs out of this graph structure to detect cell body boundary. Figure 4 shows the intermediate image outputs.

**Patch Based Data Generation:**
This approach is used to create images with real-world conditions from the high-resolution ROI regions where each image may contain multiple cervical cells along with overlaps. We use sliding window technique with stride 64 to create 128x128 patch images. The cells in the slide data are widely dispersed and contain more background. The patches containing more than 75% of background are discarded so that we get images with more cell information. The ground truth labels for the patch data are generated using the abnormal cell mask that is accurately and manually created with the help of an expert pathologist. Figure 5 shows how the abnormal cells were manually located. A patch image is labeled as abnormal if the object area in the abnormal cell mask is greater than 20% of the patch area (128x128). All the remaining patches are labeled as normal. Figure 6 shows examples of the final 128x128 labeled patch data used for the classification task.

---

**Figure 3.** ROI detection from low resolution image

**Data Generation**

We propose two methods for cell data generation from the extracted ROIs: Graph-based cell detection and patch based data generation.

**Graph-Based Cell Detection:**

We chose this method to generate one cervical cell image (Herlev-like dataset). Initially the ROI image is over-segmented by generating superpixels through QuickShift approach\(^\text{\scriptsize 17}\). The pixel intensities are averaged over each superpixel region. The resultant image is converted into a graph with centroid of superpixels as nodes, line connecting the adjacent nodes as edges and absolute difference of L2 normalized color intensities at the respective adjacent nodes as edge weight. Graph cut is used at an empirically determined threshold of 59. The resultant image is a binary mask for the nuclei present in the ROI image which form the graph nodes. The novelty of the proposed approach lies in generation of subgraphs out of this graph structure to detect cell body boundary. Figure 4 shows the intermediate image outputs. This approach could, in the future, automate segmentation of single cell images even where cells overlap.

**Figure 4.** (from left to right) Original image; Averaging pixel intensities over superpixel regions; Nuclei mask; Graph structure connecting cellular regions

**Patch Based Data Generation:**

This approach is used to create images with real-world conditions from the high-resolution ROI regions where each image may contain multiple cervical cells along with overlaps. We use sliding window technique with stride 64 to create 128x128 patch images. The cells in the slide data are widely dispersed and contain more background. The patches containing more than 75% of background are discarded so that we get images with more cell information. The ground truth labels for the patch data are generated using the abnormal cell mask that is accurately and manually created with the help of an expert pathologist. Figure 5 shows how the abnormal cells were manually located. A patch image is labeled as abnormal if the object area in the abnormal cell mask is greater than 20% of the patch area (128x128). All the remaining patches are labeled as normal. Figure 6 shows examples of the final 128x128 labeled patch data used for the classification task.
Figure 5. Manually locating abnormal cells in high resolution ROI regions.

Figure 6. 128x128 patch image data from high resolution ROIs; Class labels: Abnormal (top-row) and Normal (bottom-row).

Classification

We consider images generated from patch based approach for the binary classification task. The patch data is completely randomized and split into train, validation and test data for training different CNN models. The abnormal patches are few in number compared to the normal patches. There were 2,060 abnormal cell patches generated, which are taken into the dataset and 2,060 normal cell patches were randomly selected to be a part of the dataset. The training and validation data sets consist of the entire Herlev Pap data and a part (75%) of the patches. The training and validation data split (65-10) with remainder of the patches forming the test data. This facilitates a balanced data distribution for image classification without any bias. Table 2 shows the data distribution among training, validation and testing datasets.

Table 2. Data split for training, validating and testing CNN classifier.

<table>
<thead>
<tr>
<th>Input for CNN Classifier</th>
<th>Total</th>
<th>Patch data</th>
<th>Herlev data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>1396</td>
<td>1200</td>
<td>196</td>
</tr>
<tr>
<td>Abnormal</td>
<td>1760</td>
<td>1200</td>
<td>560</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>246</td>
<td>200</td>
<td>46</td>
</tr>
<tr>
<td>Abnormal</td>
<td>315</td>
<td>200</td>
<td>115</td>
</tr>
<tr>
<td>Testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>660</td>
<td>660</td>
<td>-</td>
</tr>
<tr>
<td>Abnormal</td>
<td>660</td>
<td>660</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>5037</td>
<td>4120</td>
<td>917</td>
</tr>
</tbody>
</table>
We used various well-established CNN classifiers in order to determine the best performing CNN model for the classification of cervical cytology patch data. Models VGG-19\(^{18}\), ResNet-50\(^{19}\), DenseNet-121\(^{20}\), and Inception_v3\(^{21}\) were fine-tuned whose weights are initialized with pre-trained ImageNet weights. In the training phase, all the layers are trained, and layer weights are updated for each epoch, and the model is trained with batch size of 32, learning rate 0.005 and momentum 0.9 (chosen empirically). This is a bi-classification task, so we employ cross entropy loss along with stochastic gradient descent optimizer. Each model is targeted to train for 500 epochs and ultimately the best weights are saved.

**Results and Discussion**

A total of 5,037 images (4,120 patch data + 917 Herlev data) were considered for the study. 1,200 patch data (660 normal, 660 abnormal) are utilized for testing and the CNN models are trained with remaining data. Pytorch deep learning platform is used to run the models on Nvidia DGX-1. The results tabulated are as shown in Table 3.

**Table 3.** Performance of various CNN models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Confusion matrix (\begin{bmatrix} TN &amp; FP \ FN &amp; TP \end{bmatrix})</th>
<th>ACC</th>
<th>PREC</th>
<th>REC</th>
<th>F1-Score</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resnet-50</td>
<td>[589 \quad 71 \quad 582 ]</td>
<td>0.8871</td>
<td>0.8913</td>
<td>0.8818</td>
<td>0.8865</td>
<td>0.7742</td>
</tr>
<tr>
<td>VGG-19</td>
<td>[581 \quad 79 \quad 592 ]</td>
<td>0.8886</td>
<td>0.8823</td>
<td>0.8970</td>
<td>0.8896</td>
<td>0.7773</td>
</tr>
<tr>
<td>DenseNet-121</td>
<td>[611 \quad 49 \quad 529 ]</td>
<td>0.8636</td>
<td>0.9152</td>
<td>0.8015</td>
<td>0.8546</td>
<td>0.7329</td>
</tr>
<tr>
<td>Inception_v3</td>
<td>[429 \quad 231 \quad 603 ]</td>
<td>0.7818</td>
<td>0.7230</td>
<td>0.9136</td>
<td>0.8072</td>
<td>0.5843</td>
</tr>
</tbody>
</table>

The results indicate that ResNet-50 and VGG-19 models performed better when compared to DenseNet-121 and Inception_v3. DenseNet-121 has 0.9152 precision (PREC) value but has poor recall (REC) of 0.8015 with 131 false negatives. On the other hand, Inception_v3 has best recall (0.9136) compared to others, but poor at precision with 231 false positives. F1-score which presents balance between precision and recall is the best measure to evaluate the model performance. The F1-score of VGG-19 being 0.8896 votes VGG-19 as better model compared to ResNet-50 (0.8865). ResNet-50 is better at precision and VGG-19 is better at recall. In the field of biomedical image analysis higher recall value is always preferable, that is, lower false negatives are always recommended. Matthews correlation coefficient (MCC) provides much balanced measure by considering true and false positives and negatives. MCC for VGG-19 at 0.7773 makes it the best model compared to ResNet-50 (0.7742). Also, accuracy (ACC) wise VGG-19 is better than ResNet-50. The relatively small differences make both the models good competitors. To better understand the performance of the models at various classification thresholds, we plot the receiver operating characteristic (ROC) curve. Figure 7 shows the ROC plots for all the four models.

The best operating Q-point for the models in the cervical cytology image classification is the point where the curve has high sensitivity (true positive rate) and high specificity (1 – false positive rate). VGG-19 and ResNet-50 performed similarly with their best Q-points compared to other models. The accuracy under the ROC curve (AUC) was found to be 0.95 for both the models.

ResNet-50 and VGG-19 are good at generalizing the cervical cytology data with the pre-initialized ImageNet weights. VGG-19 is a shallow network compared to models under study and uses 3x3 convolutional layers stacked up along the depth of the network has proven to be a better model for the cervical cytology classification task in our study. ResNet-50 being a 50-layer deeper network, addresses its vanishing gradient problem through their residual learning blocks. This makes it a good competitor against VGG-19. Each layer in DenseNet-121 is fed with the outputs from the previous layers which improve the feature propagation and alleviate the vanishing gradient problem. Although DenseNet-121 is more efficient on some image classification tasks, it could not outperform the VGG-19 model on the cytology image classification. The 42-layer deep Inception v3 model factorizes convolutions and aggressively reduces dimensions which reduces the computational cost but could not maintain the quality in classifying the cytology image data.
Figure 7. ROC curve of the four CNN models

We have further explored the performance of our best model, VGG-19, on the 128x128 abnormal cell patch data extracted with respect to the centroid of each object in the abnormal mask. 203 abnormal image patches were obtained and tested with VGG-19 model. We achieved an accuracy of 0.8778 upon prediction on these 203 abnormal image patches. This indicates that the VGG-19 model is better at classifying the cytology cell images, even under challenging cell-overlapping conditions.

While this work is preliminary, it demonstrates the capability of deep learning to recognize abnormal cells in cervical cytology specimens. With further refinement this work could be incorporated into a production level tool to assist pathologists with pre-screening and quality assurance, thereby improving a pathologist’s efficiency and accuracy.

Conclusion

We have successfully developed and evaluated a prototype pipeline for the classification of cervical cytology slide images. The process automatically generated cleaner labeled patch image data for training and testing convolution neural networks. Our approach considers realistic conditions of overlapping cells which is superior to state-of-the-art classification techniques that rely on segmented cells. We investigated various CNN models for successful classification of cytology image data, and found VGG19 and ResNet-50 were similar best performers with our data. A novel graph-based cell detection technique was also proposed which may be used for developing cell analysis techniques. Our work represents a novel approach for classifying cytopathology image data, using real-world samples.

Acknowledgment

This research is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications. In addition, we thank Becton Dickinson Diagnostic Services for de-identified SurePath liquid cytology Pap slides and Stephen Hewitt, MD, PhD in the Laboratory of Pathology, National Cancer Institute for assistance with whole-slide scanning.

References

Why Patient Portal Messages Indicate Risk of Readmission for Patients with Ischemic Heart Disease

Lina Sulieman, PhD, Zhijun Yin, PhD, Bradley A. Malin, PhD
1 Vanderbilt University, Nashville, TN

Abstract

Online portals enable patients to exchanging messages with healthcare providers. After discharge, patients message providers to ask questions and report problems. Care providers read and respond accordingly, which requires a non-trivial amount of human effort and is unlikely to scale up as portals become more popular. Automatically detecting when a message indicates a worsening in a patient’s condition can assist providers to identify patients at risk of readmission. We investigated the association between messages that patients, diagnosed with ischemic heart disease, sent after discharge and the risk of readmission. We studied 4,052 messages sent after discharge for 1,552 patients. We represented messages using inferred latent topics, linguistic features (e.g. emotions, activities), and clusters of medical terms. Our analysis indicates that mentioning medication dosage and additional procedures are associated with readmission. Moreover, patients who were readmitted rarely mentioned leisurely activities or described their insights about their health information.

Introduction

Patient portals are secure online websites that healthcare organization provide to grant patients 24 hour access to their health records. Portals include a wide range of health information, including discharge summaries, medications, immunizations, and laboratory tests. One of the popular functionalities of patient portals is their support of secure messaging between patients and care providers. Messaging allows patients to keep their healthcare providers informed about their clinical status outside of visits to the clinic and stays in the hospital. Moreover, The use of messaging has been shown to be associated with improved chronic disease management and medication adherence. It has been shown that discussions about laboratory test results, reporting new symptoms, and requesting prescription refills are the most common topics in patients’ messages. The popularity of these topics demonstrates that patients utilize the portal messages for seeking information from healthcare providers when they are outside of the clinical environment. In addition, patients communicate different needs in their messages, which can be roughly partitioned into logistical (e.g., location of clinic), social (e.g., thanking care providers), informational (e.g., asking about intervention), and medical (e.g., informing physician of a health problem). Notably, it has been shown that more than 70% of messages originating from patients included medical needs.

There has been limited research into the identification of the symptoms and events that patients communicate through their messages, and their relationship with outcomes such as medication discontinuation or readmission. This may be due, in part, to the fact that the information relevant to a patient’s clinical status in a message is not explicitly indicated, but rather is in an unstructured form. This makes it essential to first identify potential factors that are indicative of health issues in portal messages, so that healthcare providers can assess and evaluate the health status of a patient. This need is further exacerbated by the rapid growth in the adoption of this technology and the number of patient messages as an artifact. Hence, healthcare organizations would benefit from automating the process of detecting a message that can indicate the risk of a negative outcome, such as readmission.

In this study, we aim to identify the contents and textual features in patient messages that can indicate the possibility of negative outcomes. Specifically, we analyzed the content of messages sent by patients who were diagnosed with ischemic heart disease and hospitalized at Vanderbilt University Medical Center (VUMC). We analyzed the content of messages sent after discharge and within 31 days for two types of patients: 1) those who were readmitted due to an unplanned hospitalization and 2) those who lacked an unplanned hospitalization. We extracted linguistic and textual features, along with patients’ demographics, and applied a generalized linear model to learn their association with readmission risk.

Related work

Several studies have analyzed the content of messages sent through patient portals. Some of these investigations relied upon manually review, while others applied machine learning and statistical analysis to automatically extract patients’ needs and assess the association between the messages and an event of interest. Certain studies focused on the volume or content of messages and with respect to outcome. For instance, Sulieman et. al. investigated the post-
discharge factors that are associated with readmission risk\textsuperscript{19}. They found that the number of messages that patients sent after discharge was one of the top predictors of readmission. Yin et. al. extracted the patterns of messaging with healthcare providers, the volume of messages and the content of messages sent by breast cancer patients\textsuperscript{15}. The authors combined these features to find associations between messages and the potential for discontinuing hormonal therapy\textsuperscript{15}. They observed that mentions of side effects and surgery-related topics were associated with an increased risk of discontinuation. By contrast, they further observed that expressions of gratitude and mentions of drugs prescribed to treat side effects were associated with a decreased risk of discontinuation. North et. al. reviewed and assessed the content of patient messages and its association with the risk of death within 30 days and the risk of hospitalization within 7 days following the message\textsuperscript{17}. They found that patients mentioned high risk symptoms in 3.5% of messages and that six hospitalizations (0.09% of messages) were related to a patient message.

**Methods**

**Cohort**

We extracted data from the VUMC Synthetic Derivative (SD), a de-identified version of the electronic health record (EHR). We focused our analysis on patients who exhibited ischemic heart disease during an inpatient visit and sent a message through the MyHealthAtVanderbilt (MHAV) patient portal after discharge. We identified the patients who were readmitted within 31 days. We excluded patients who did not send any messages after discharge. We also retrieved patients’ age at discharge, gender, race, and ethnicity.

There were 96,044 patients who were diagnosed with ischemic heart disease, with admissions between 1990 and 2018. 6,448 of these patients sent a message using MHAV between 2003 and 2018. The cohort for this study consisted of the 1,552 patients who sent a message within 31 days after the discharge. As shown in Table 2, approximately two-thirds of the patients were male and with an average age of 63. The patients were 93% Caucasian and 98% non-Hispanic.

<table>
<thead>
<tr>
<th>Demographic Feature</th>
<th>Without readmission N = 1512</th>
<th>Readmitted N = 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>63.0 (12.3%)</td>
<td>62.1 (15.8%)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1024 (67.7%)</td>
<td>25 (62.5%)</td>
</tr>
<tr>
<td>Female</td>
<td>488 (32.3%)</td>
<td>15 (37.5%)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1408 (93.1%)</td>
<td>35 (87.5%)</td>
</tr>
<tr>
<td>Black</td>
<td>79 (5.2%)</td>
<td>5 (12.5%)</td>
</tr>
<tr>
<td>Asian</td>
<td>13 (0.9%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Native American</td>
<td>3 (0.2%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>9 (0.6%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NonHispanic</td>
<td>1487 (98.3%)</td>
<td>39 (97.5%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>16 (1.1%)</td>
<td>2 (2.5%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>9 (0.6%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

**Message Extraction and Analysis**

We extracted the messages exchanged between patients and their healthcare providers, regardless of who initiated the message. For patients who were readmitted, we extracted the messages they sent after discharge and before readmission. For patients without a readmission, we extracted the messages sent within 31 days after the discharge. For patients with multiple admissions, we included the earliest readmission event only.
Text analysis

We grouped the messages sent by each patient after their discharge into a single document. To profile the content of each patient’s message, we preprocessed the messages, performed topic analysis, extracted linguistic features, and defined medical term clusters, as outlined in Figure 1.

Topic Analysis

To extract topics from messages, we applied Latent Dirichlet Allocation (LDA) as implemented in the Mallet Java package (version 2.0.8). LDA is a popular topic modeling method that allows a group of documents to be explained by latent topics, each of which can be further explained by the words in the documents. After running LDA, we obtained a topic distribution for each document (e.g., the probability that a document can be explained by each topic), and a word distribution for each topic (e.g., the probability that a topic can be explained by each word). Based on these distributions, together with the lengths of the documents, we calculated the topic distribution across the corpus by combining all the documents into a single large document. LDA has proven effective at summarizing a large amount of text. Since it is an unsupervised machine learning method, we relied on the coherence score to determine the best number of topics. The coherence score is used to measure the extent to which the most probable words in every topic appear together in either the current documents or some external data source (e.g., Wikipedia). A higher coherence score suggests a better topic modeling result. We learned LDA models with 2 to 26 topics (with a step size of 1) and chose the number of topics that exhibit the largest coherence score. To mitigate word sparsity and ensure interpretability, we replaced each term with its lemma form and retained only nouns, verbs, adjectives and adverbs. We also generated the bi-grams of terms using the genism python package (version 3.6.0) to capture more meaningful phrases.

Linguistic Features

We applied Linguistic Inquiry and Word Count (LIWC, version 2015) to extract the cognitive, emotional, and social aspects in the messages. LIWC is an effective tool to summarize linguistic features from online generated content. The LIWC package generates approximately 90 linguistic variables, including general descriptor categories (e.g., words per sentence), standard linguistic dimensions (e.g., percentage of pronouns in the message), word categories tapping psychological constructs (e.g., affect, cognition, biological processes, and drives), personal concern categories (e.g., leisure, work, and home), informal language markers (e.g., assents and swear words), and punctuation categories. In our text analysis, we focused on psychological constructs, personal concerns, and informal language. Table 1 shows the 45 features that LIWC extracted.
readmission was

From the figure, readmi

The patients sent 4

Results

statistical models

associations, and the

For each model, we identified the features with coefficient

45 linguistic features, and 35 word2vec clusters

We applied

Content

the elbow principle to locate the cluster number where the marginal gain of increasing cluster size

tends to become small as the number of clusters increases.

We applied

Medical Term Clusters

We used Clamp (version 1.5.0) to extract the medical terms from the messages, including treatments, clinical problems, and laboratory tests. Given the large number of clinical terms that were extracted, we reduced the dimensionality to represent features more efficiently by grouping terms that shared similar meanings. To do so, we first trained a word2vec model using the clinical communications in the entire SD. We did not use the pretrained word2vec model (e.g., the Google word2vec) because there are many terms (e.g., abbreviations) that do not exist in the pretrained documents. Word2vec generates a vector (or embedding) for each word, where similar words exhibit high semantic similarity based on a cosine function. We trained word2vec using the genism python package with a minimum word count of 50, window size of 15, and 100 hidden units.

We retrieved the words’ vectors (i.e., word2vec embedding) for each medical term in the messages extracted by Clamp. For the cases where a medical term consists of more than one word, we retrieved the word2vec embeddings for each word and calculated the mean. We applied agglomerative hierarchical clustering with complete linkage, according to a cosine distance, in the sklearn python package (version 0.20.0) to cluster the word embeddings. To obtain the most efficient number of clusters, we adopted a metric that generates a number of clusters that is 1) large enough to create efficient and interpretable semantic clusters and 2) small enough to avoid partitioning one cluster into two or more clusters with similar words. This was accomplished through the approach introduced by Yin et. al. Specifically, we construct clusters where the number of clusters range from 2 to 100 clusters (with a step size of 1). We use the standard deviation of the cluster sizes to determine when to stop the clustering process. Heuristically, the standard deviation tends to become small as the number of clusters increases. To identify the optimal number of clusters, we followed the elbow principle to locate the cluster number where the marginal gain of increasing cluster size begins to diminish.

Content Analysis

We applied logistic regression to evaluate the associations between the content of messages that includes 19 topics, 45 linguistic features, and 35 word2vec clusters and demographics, and the readmission risk. Specifically, we used the Generalized Linear Model (GLM) library in R (version 3.5.2) to learn three association models:

1- Demographics-Only model: This includes age (rescaled to 0-1 range), gender, race, and ethnicity;
2- Message content model: The learned topics (rescaled to [0,1] range), LIWC linguistic features (rescaled to [0,1] range), and medical term clusters; and
3- Demographics and message content model: A combination of models 1 and 2.

For each model, we identified the features with coefficients that were statistically significant at the 0.05 level, their associations, and the Akaike information criterion (AIC) values. The latter is an estimate of the relative quality of statistical models, which is obtained by estimating the amount of information loss in the model. A higher AIC value indicates a lower amount of information loss and thus a better quality.

Results

The patients sent 4,052 messages either before readmission or within 31 days after the discharge (for those without a readmission). Figure 2 illustrates the distribution of messages sent by patients according to their readmission status. From the figure, it can be seen that around 60% and 80% of the patients who were readmitted and lacked a readmission, respectively, sent only one message. The average (median) number of messages sent by patients with and without a readmission was 1.75 (1) and 1.87 (1), respectively. We ran Mann-Whitney to evaluate whether the numbers (e.g.,
number of messages, number of words in messages) were statistically different for patients with readmission and patients without readmission. The difference between the number of messages sent by those two groups was not found to be statistically significant (Mann-Whitney U: statistic=28120, p-value = 0.2). Among the readmitted patients, 20% sent two messages, while only 10% of the non-readmitted patients sent two messages. On average, the messages sent by patients consisted of 700 words overall (median = 381 words), while the readmitted and non-readmitted subgroups consisted of 667 words (median = 546 words) and 700 words (median = 376 words), respectively. The difference between the number of words in messages from patients with and without a readmission was not found to be statistically significant either (Mann-Whitney U: statistic = 27291.0, p-value = 0.15).

More than 50% of readmitted patients sent a message within the first five days after discharge, while 30% of patients who were not readmitted sent a message within 5 days, as depicted in Figures 2(b) and 3(a). Only 10% of patients sent a message five days before the readmission. The mean of the day of the first message sent by a readmitted and non-readmitted patient was 7.8 (SD = 6.3 and median = 5) and 12.9 (SD = 9 and median = 12), respectively. The difference between the day of the first message sent by readmitted patients and non-readmitted patients was statistically significant (Mann-Whitney U: statistic = 20478.0, p-value = 0.002).

![Bar chart](image1.png)

(a) Patients not readmitted within 30 days  
(b) Patients readmitted with 30 days

*Figure 2.* Number of messages sent post-discharge sent by patients (a) without and (b) with a readmission.

![Bar chart](image2.png)

(a) Patients not readmitted within 30 days  
(b) Patients readmitted with 30 days

*Figure 3.* First day a message was sent after discharge by a patient (a) without and (b) with a readmission.
Message Topics

We identified 19 topics in post-discharge messages. Each topic includes a set of words that patients invoked to discuss a particular topic. For instance, when patients ask about, or reschedule, appointments, they use relative phrases such as the day of the week, time, schedule, confirm, or reschedule. Table 3 lists the most relevant words in each topic ranked by the LDA model. It can be seen that the topics primarily covered appointments, vitals, checking laboratory tests, medications (including prescription and time), and logistics (including discharge locations and communications).

Readmission Risk Associations

The demographic model achieved an AIC of 379; however, none of the features (i.e., age, gender, race, and ethnicity) were statistically significant. In the message content model, we applied 19 topics, 45 linguistic features, and 35 clusters. This model achieved higher quality with an AIC of 462 and contained six features that had a statistically significant association with the readmission risk (as shown in Table 4). We report the statistically significant features. The swear linguistic category was significant and positively correlated with the readmission event. By contrast, both insight and leisure were negatively correlated with the readmission event. Moreover, the words in three clusters listed in Table 5 were significantly associated with readmission. Each cluster includes the words that are similar to each other based on their Word2Vec similarity scores. The words in Clusters 2 and 29 were positively correlated with readmission, while those in Cluster 11 were negatively correlated. It should be noted that removing correlated values from the model did affect the significance of the features. When combining demographic features and content features, we obtained a model with an AIC of 468, which was slightly higher than the message content model. In this combined model, only one feature, Cluster 2 was significant. The demographic features were still insignificant in this model.

Table 5 shows the words for each significant word semantic cluster: Cluster 2, Cluster 11, and Cluster 29. We ranked the words in each cluster based on their cosine distance to its centroid (i.e., the mean of the word2vec for the words in the cluster). As Table 5 shows, the most relevant words in Cluster 2 primarily correspond to the medication dosage that was prescribed for the patient. The top words in Cluster 11 are laboratory test names, while Cluster 29 includes the mentions of another clinical event, such as a procedure, laboratory test, and medication.

Table 3. The 30 most relevant words in each of the topics extracted from the messages.

<table>
<thead>
<tr>
<th>#Topic</th>
<th>Topic</th>
<th>30 Most Relevant Terms</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Appointments</td>
<td>appointment, schedule, appt, dr, follow, week, reschedule, time, hospital, cancel, make, friday, nov, tuesday, clinic, thursday, apr, office, jan, feb, work, aug, advise, discharge, pm, early, confirm, tomorrow, monday, afternoon</td>
<td>6.53%</td>
</tr>
<tr>
<td>2</td>
<td>Monitoring</td>
<td>weight, day, lasix, morning, today, continue, fluid, send, blood, pressure, potassium, leg, swell, breath, week, daily, increase, feel, foot, advise, dose, lb, shortness, pill, time, afternoon, bumex, mg, yesterday, walk</td>
<td>5.89%</td>
</tr>
<tr>
<td>3</td>
<td>Vital (specifically blood pressure)</td>
<td>blood, low, pressure, heart, start, send, heart, rate, pulse, feel, record, dr, week, check, normal, back, review, time, monitor, rn, amiodarone, med, rate, make, home, nr, office, episode, today, medication, drop</td>
<td>5.87%</td>
</tr>
<tr>
<td>4</td>
<td>Laboratory test checks</td>
<td>lab, result, order, draw, blood, week, level, work, clinic, friday, tomorrow, repeat, check, test, dr, day, good, wait, nr, bmp, today, iron, back, home, dose, potassium, low, recheck, send, cbc</td>
<td>5.80%</td>
</tr>
<tr>
<td>5</td>
<td>Time specifically for medication</td>
<td>mg, daily, tablet, day, medication, increase, med, change, metoprolol, hour, list, hospital, morning, dose, mouth, continue, question, review, make, feb, pressure, follow, dosage, week, tab, add, bid, furosemide, procedure, start</td>
<td>5.77%</td>
</tr>
<tr>
<td>6</td>
<td>Feeling at time</td>
<td>good, feel, today, hope, morning, great, glad, make, check, talk, tomorrow, yesterday, increase, hospital, start, thing, weekend, week, hear, night, home, feeling, bit, weak, bsn, change, ve, time, wonderful, long</td>
<td>5.48%</td>
</tr>
<tr>
<td>7</td>
<td>Communications</td>
<td>call, phone, starrser, patient, dr, message, number, back, speak, leave, today, pt, nurse, request, reach, rn, wife, state, return, give, regard, cell, yesterday, notify, result, response, md, provider, miss, set</td>
<td>5.45%</td>
</tr>
<tr>
<td>8</td>
<td>Discomfort feelings and symptoms</td>
<td>pain, chest, good, leg, time, side, continue, area, leave, incision, problem, arm, normal, walk, due, ms, tylemol, bad, hurt, feel, discomfort, back, drainage, improve, dr, clear, hip, heal, put, level</td>
<td>5.25%</td>
</tr>
<tr>
<td>9</td>
<td>Tests</td>
<td>blood, test, problem, heart, time, stent, case, cath, cardiologist, symptom, cardiac, month, prior, urine, stress, week, back, result, stress, t, recommend, risk, plan, good, feel, study, put, show, remember, year, plavix</td>
<td>5.25%</td>
</tr>
</tbody>
</table>
### Table 4.
Features with statistically significant beta coefficients in the GLM model based on message content.

| Concept                | Estimate | Std. Error | z-value | Pr(>|z|) |
|------------------------|----------|------------|---------|----------|
| swear                  | 7.375    | 3.139      | 2.350   | 0.0188   |
| insight                | -0.655   | 0.317      | -2.066  | 0.0389   |
| leisure                | -1.893   | 0.963      | -1.966  | 0.0493   |
| Cluster 2: Medication dosage | 1.196    | 0.503      | 2.377   | 0.0175   |
| Cluster 11: Laboratory test | -2.963   | 1.257      | -2.357  | 0.0184   |
| Cluster 29: “Another” event | 1.404    | 0.669      | 2.098   | 0.0359   |

### Table 5.
The top 20 words in each statistically significant cluster. The words are ranked according to their distance from the centroid of the cluster in ascending order.

<table>
<thead>
<tr>
<th>Cluster 2: Medication dosage</th>
<th>Cluster 11: Laboratory test</th>
<th>Cluster 29: “Another” event</th>
</tr>
</thead>
<tbody>
<tr>
<td>prednisone dosage reduction</td>
<td>labs (cpd cmp ldh)</td>
<td>another cxr</td>
</tr>
<tr>
<td>current dose diabetes medication</td>
<td>cbc cmp</td>
<td>another mri</td>
</tr>
<tr>
<td>higher dose steroids</td>
<td>cbc cmp</td>
<td>another bmp</td>
</tr>
<tr>
<td>usual gabapentin dose</td>
<td>cdp cmp ldh uab</td>
<td>another echo</td>
</tr>
<tr>
<td>tavr procedure dose</td>
<td>cdp cmp ldh</td>
<td>another dilemm</td>
</tr>
<tr>
<td>prednisone dosage</td>
<td>cmp cbc bnp</td>
<td>another treatment</td>
</tr>
<tr>
<td>methotrexate dose</td>
<td>cdp cmp igg igm spep</td>
<td>another procedure</td>
</tr>
<tr>
<td>lortab dosing</td>
<td>cdp cmp</td>
<td>another paracentesis</td>
</tr>
</tbody>
</table>
nebulizer meds  cpd cmp aml  another uti
shot oral dose steroids  cmp cpd  another ultrasound
current prednisone dose  cpd cmp gengraf level  another medication
vancomycin dose  lipids cmp  another ct scan
chemo treatment  lipids cmp  another ct-scan
radio frequency ablation procedure  cpd cmp igg spep serum free light chains code  another cbc
normal asacol dose  cbc cmp ps type screen  another diuretic
current medication regimen  fasting lipids cmp  another ct
paxil dose  cmp ldh  another xray
maintenance dose prednisone  bmp  another infection
prednisone dose  cbc+diff bmp  another antibiotic

Discussion

This investigation yielded several notable findings. First, the message patterns for patients who were readmitted were different from patients who were not readmitted. Specifically, patients who were readmitted tended to send messages earlier than other patients. This might indicate that these patients were experiencing problems or complications after discharge that they communicate to their healthcare providers. Identifying the relevant words, topics, or signals in the message may assist healthcare organizations to identify patients who are at higher readmission risk and, thus, address such complications in a timely manner.

Second, patients’ messages included indications regarding patients’ health status, health concerns, and social context after discharge, which may be useful for predicting readmission. While conventional features, such as patient demographics did not exhibit a significant association with the readmission status, the model performance was improved by incorporating the message content. We believe this is because it includes patients’ activities and concerns. For instance, social information that patients communicated about their leisurely activities, insights, and feelings were statistically significant. The analysis demonstrated that readmitted patients were less likely to write sentences about leisure activities or describe patients’ intuitions and insights (e.g., think or know) as the coefficients of the model indicated. For example, patients who were not readmitted tended to send similar questions, seek feedback or answers, such as “I think my main concern is how the H/H was trending down at the time of discharge”. One of the messages exchanged between a provider and a patient included a mention of hunting, walking (e.g., “gone deer hunting”, “I walk about half a mile to get my hunting school”, “restricted to walking a half hour daily instead of an hour”). This suggests that the readmitted patients might be experiencing unfamiliar symptoms or feelings, thus limiting their ability to enjoy some social or leisurely activities. Moreover, readmitted patients tended to use swear words in their messages, which may indicate their frustration after discharge.

Third, the medical terms in the model had a significant association with the status of the patient after discharge. Mainly, the messages that have information about the dosage of patient medications and the laboratory tests were significantly associated with readmission. These medical information types exhibited opposing directionality in their association with a patient’s outcome. Specifically, the medication dosage had a positive association with readmission, which might indicate that the readmitted patient asked about their medication dosages or explained some side effects with the prescribed dosage. By contrast, the existence of laboratory values in the message had a negative association with readmission. Another significant medical feature is the request to repeat or perform another clinical test or another medication. For example, the mention of an additional clinical event such as another MRI, another CT scan, another UTI have a positive association with readmission. Hence, the necessity of ordering another scan or another test can imply that the patient had complications after the discharge, which increased the probability of readmission.

At the same time, there are several limitations worth noting. First, this a study of a specific population at a single medical center, which calls into question the generalizability of our findings. Second, the dataset size was relatively small. Expanding this analysis to include more phenotypes and a large number of messages could provide more intuition into the associations. Third, we combined the messages sent by a patient, which masks the temporal changes in the topics. Fourth, our text analysis did not handle the negation which we will address in our future work. Fifth, our analysis focused on identifying the indication of readmission in patients messages. We did not evaluate the capability of predicting the readmission using the model. In our future work, we will focus on evaluating the ability of predicting the readmission using the model. Finally, there was a low readmission rate in our dataset. Sending MHAV secure
messaging might indicate that the patients used online patient portal to seek advice/information regarding their health, which might explain, in certain degree, the low rate of readmission in MHAV cohort who seeks information in the portal.

**Conclusion**

Online portals provide a secure channel that allows patients to interact with their healthcare providers. Patients use portal messages to communicate their needs, requests, and questions. While the number of patient portal messages is increasing, analyzing the patient reported information in their messages is still limited. Identifying the signals in portal messages that indicate the risk of readmission can help providers apply interventions to avoid adverse events. This study showed the messages sent by patients with ischemic heart disease after a hospital discharge can be leveraged to predict readmission. The findings specifically showed that leisure activities, intuition (e.g., think), and swear words, as well as medical terms in the messages are associated with readmission events. We believe that future research will benefit by evaluating the capability of our model to predict the readmission using patients message and expanding on the analysis to include other diseases.

**Acknowledgements**

This research was sponsored in part by NIH grant UL1TR002243.

**References**

Integrating Clinical Knowledge and Real-World Evidence for Type 2 Diabetes Treatment

Xingzhi Sun, PhD\textsuperscript{1}, Wei Zhao, PhD\textsuperscript{1}, Lei Zuo, MS\textsuperscript{1}, Alexandra Dumitriu, PhD\textsuperscript{2}, Chuang-Chung Lee, PhD\textsuperscript{2}, Nan Cui, MD\textsuperscript{3}, Xiyang Liao, MS\textsuperscript{1}, Tingting Zhao, MS\textsuperscript{1}, Xuehan Jiang, PhD\textsuperscript{1}, Zhuoyang Xu, MS\textsuperscript{1}, Gang Hu, PhD\textsuperscript{1}, Guotong Xie, PhD\textsuperscript{1}, Hong Wu, MD\textsuperscript{3}, Yahua Huang, MS\textsuperscript{3}

\textsuperscript{1}Ping An Health Technology, Beijing, China; \textsuperscript{2}Sanofi, Cambridge, Massachusetts, U.S.A.; \textsuperscript{3}Sanofi (China) Investment, Shanghai, China

Abstract

Clinical decision support system (CDSS) plays a significant role nowadays and it assists physicians in making decisions for treatment. Generally based on clinical guideline, the principles of the recommendation are provided and may suggest several candidate medications for similar patient group with certain clinical conditions. However, it is challenging to prioritize these candidates and even refine the guideline to a finer level for patient-specific recommendation. Here we propose a method and system to integrate the clinical knowledge and real-world evidence (RWE) for type 2 diabetes treatment, to enable both standardized and personalized medication recommendation. The RWE is generated by medication effectiveness analysis and subgroup analysis. The knowledge model has been verified by clinical experts from the advanced hospitals. The data verification results show that the medications that are consistent with the method recommendation can lead to better clinical outcome in terms of glycemic control, compared to those inconsistent.

Introduction

Type 2 diabetes (T2DM) is a chronic disease that often leads to different types of complications and causes serious health problems. According to the eighth edition of International Diabetes Federation (IDF) diabetes atlas\textsuperscript{1}, in 2017, China has more than 114 million diabetes patients, accounting for 27\% of the whole diabetes population worldwide. In addition to this challenging situation, China faces severe shortage of experienced physicians, especially in the large rural area. According to China Statistical Yearbook (2018)\textsuperscript{2}, 41\% of Chinese population lives in the rural area, but only consumes 21\% of national medical resources. As stated in China Health and Family Planning Statistical Yearbook (2017)\textsuperscript{3}, comparisons of health workers were observed between rural and urban areas for health professionals (4.0 vs. 10.8), doctors (1.6 vs. 4.0), Registered Nurses (RNs) (1.5 vs. 5.0) per 1,000 population. In this background, a robust clinical decision support system (CDSS) for type 2 diabetes treatment is of high importance in China’s market. Given the T2DM patient’s information, the CDSS can recommend the medications and provide the evidence of why the recommendation is made. It not only assists inexperienced physicians in prescribing the appropriate medications for T2DM patients, but serves as a tool for physician education.

CDSSs have been widely applied for reducing medical errors and increasing health care quality and efficiency\textsuperscript{4,5}. There are two types of approach for CDSS: knowledge-driven and data-driven. In the context of medication recommendation, the knowledge-driven approach is to develop an expert system based on clinical knowledge (such as clinical guidelines and consensus)\textsuperscript{6,7,8}, while the data-driven one is to apply the data mining techniques on electronic health records (EHRs) to build the mapping between patient’s information and medications\textsuperscript{9,10,11}.

The advantage of knowledge-driven approach is that the recommendations are always consistent with the guidelines, which is critical to a CDSS. However, the limitation is that the knowledge provided in clinical guideline could be too general to provide fine-granular personalized recommendation\textsuperscript{12}. For example, an if-then rule for knowledge-driven CDSS could be: if one has been taking biguanides (MET) for about 3 months and his HbA1c is between 7\% to 9\%, then he needs to switch to dual therapy by adding another type of oral antidiabetic drug (OAD). However, the OAD includes a number of drug classes such as sulfonylureas (SU), glinides (GLN), and alpha-glucosidase inhibitors (AGIs) etc. The physician may still have the problem to decide which drug, often a product name in a drug class hierarchy, to be prescribed.

With the large amounts of EHRs, data-driven CDSS can be developed to provide personalized medication recommendation. Recently, Liu et al. proposed an algorithm to group the patients based on the similarity metric learnt from the real-world clinical data\textsuperscript{13}. Likewise, Chen and Altman reported a Bayesian conditional probability model for
recommendation of clinical orders through the data mining of EHRs\textsuperscript{10,11}. Although data-driven approach provides personalized result, the recommended results could violate the guideline, which degrades the physicians’ trust to use the system.

In this study, we propose a method to integrate the clinical knowledge and real-world evidence (RWE), for enabling the standardized (i.e., consistent with knowledge) and personalized (i.e., referring to the RWE of similar patients) medication recommendation for T2DM patients. In our approach, the clinical guideline provides the principles of the recommendation\textsuperscript{14} and may suggest several candidate drugs/drug combinations for the patients with certain clinical conditions. The RWE is generated for prioritizing these candidates and even refining the guideline to a finer level. Specifically, we represent the clinical guideline as a decision tree, in which each leaf node is a decision point with recommended modifications and the path from the root to the leaf node defines the clinical conditions of the corresponding subgroup of T2DM patients. Then the medical data are fed via the root of tree so that each leaf node is associated with a set of data samples. For each leaf node with multiple medication options, we perform the drug effectiveness analysis\textsuperscript{15} and subgroup analysis\textsuperscript{16} on the data samples, and associate the results as the evidence. These results of the leaf node can be applied to prioritize the corresponding candidate medications, or once confirmed by domain experts, to refine the clinical guideline, i.e., to further split the leaf node into a sub-tree. We have realized such an approach and developed a CDSS for T2DM treatment. We build our knowledge model based on the national clinical guideline for T2DM 2017\textsuperscript{17} and integrate it with RWE extracted from 119236 records of T2DM patients in a Chinese city. The knowledge model has been verified by clinical experts from the advanced hospitals.

Methods

Knowledge model for standardized medication

The knowledge model was built based on Chinese guideline for prevention and treatment of type 2 diabetes (2017), and had been verified by clinical experts from the advanced hospitals. An extracted rule contains information of related feature conditions and corresponding medication recommendation. According to the guideline, the glycemic control target could be set as the glycosylated hemoglobin (HbA1c) goal of less than 7.0% or fasting plasma glucose (FPG) goal of less than 7.0 mmol/L. If the target is not achieved without any therapy or after about 3 months of therapy, it should proceed to clinical treatment or adjustment of original treatment. So the treatment could be suggested by current blood glucose level, previous treatment and its duration.

Table 1. Extracted rules from guideline for type 2 diabetes

<table>
<thead>
<tr>
<th>Status</th>
<th>Blood glucose level</th>
<th>Previous treatment</th>
<th>Medication recommendation</th>
<th>GroupID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untreated</td>
<td>HbA1c &lt; 7.0% and FPG&lt;7.0mmol/L</td>
<td>/</td>
<td>Lifestyle management</td>
<td>Lifestyle Management</td>
</tr>
<tr>
<td></td>
<td>(HbA1c &gt;= 7.0% and HbA1c &lt; 9.0%) or (FPG&gt;=7.0mmol/L and FPG&lt;11.1mmol/L)</td>
<td>/</td>
<td>MET</td>
<td>AGIs</td>
</tr>
<tr>
<td></td>
<td>HbA1c&gt;=9.0% or FPG&gt;=11.1mmol/L</td>
<td>/</td>
<td>basal_insulin+prandial_insulin</td>
<td>premixed_insulin</td>
</tr>
<tr>
<td>Treated</td>
<td>HbA1c&lt;7.0%and FPG&lt;7.0mmol/L</td>
<td>/</td>
<td>Continued access to this therapy</td>
<td>Continued Therapy</td>
</tr>
<tr>
<td></td>
<td>HbA1c&gt;=7.0% or FPG&gt;=7.0mmol/L</td>
<td>Single drug, no insulin</td>
<td>Two drug categories with different mechanisms (with or without basal insulin)</td>
<td>Dual Therapy</td>
</tr>
<tr>
<td></td>
<td>HbA1c&gt;=7.0% or FPG&gt;=7.0mmol/L</td>
<td>Double drug, no insulin</td>
<td>Three drug categories with different mechanisms (with or without basal insulin)</td>
<td>Triple Therapy</td>
</tr>
<tr>
<td></td>
<td>HbA1c&gt;=7.0% or FPG&gt;=7.0mmol/L</td>
<td>Three drugs or use insulin</td>
<td>Three drug categories with different mechanisms (with or without basal insulin)</td>
<td>Combination Injectable Therapy</td>
</tr>
</tbody>
</table>


Drugs commonly used for type 2 diabetes treatment include seven categories of OAD: biguanides (MET), sulfonylureas (SU), glinides (GLN), alpha-glucosidase inhibitors (AGIs), thiazolidinediones (TDZs), glucagon-like peptide-1 (GLP-1) and dipeptidyl peptidase IV (DPP-4). Since SU and GLN have the same cellular mechanism called insulin secretagogues, they can be regarded as one category of “SU/GLN”.

For each extracted rule for treatment, we set a group ID according to its corresponding therapy based on knowledge model, including Lifestyle Management, Monotherapy, Short-term Intensive Insulin Therapy, Continued Therapy, Dual Therapy, Triple Therapy, Combination Injectable Therapy. Each group contains several treatment options, as shown in Table 1.

For example, information like “if lifestyle alone fails to achieve blood glucose control, monotherapy should be initiated” in guideline would be converted to a rule like: if \( \text{HbA1c} \geq 7.0\% \) and \( \text{HbA1c} < 9.0\% \) or \( \text{FPG} \geq 7.0\,\text{mmol/L} \) and \( \text{FPG} < 11.1\,\text{mmol/L} \), then initiate Monotherapy (MET or SU/GLN or AGIs). Metformin (MET) should be the first priority when initiating therapy.

**Real-world evidence integration for personalized medication**

Based on the knowledge model, the clinical guideline provides the principles of the recommendation and may suggest several candidate drugs/drug combinations for the patients with certain clinical conditions. The RWE is generated from real-world data to suggest the priority of these candidates and even refine the guideline to a finer level. Evidences mainly include feature statistics that describe a group of similar patients, usage percentage of candidate medications in the patient group and their corresponding outcome of glycemic control. When physicians face several options, RWE data (e.g. outcome or effect evaluation) become very important for identifying the optimal and personalized treatment. Figure 1 shows our methodology of integrating clinical knowledge and real-world evidence for type 2 diabetes treatment. First, we represent the knowledge model derived from clinical guideline as a decision tree, in which each leaf node is a decision point (corresponding to each GroupID) with recommended modifications and the path from the root to the leaf node defines the clinical conditions of the corresponding group of T2DM patients. In step 1 of Figure 1, the medical data are fed via the root of tree so that each leaf node (GroupID) is associated with a set of data samples. In step 2, by knowledge model, the data samples could be divided into two parts of guideline-concordant and guideline-not-concordant samples depending on its prescription, and we perform data-based evaluation of standardized medication and its relationship with effectiveness. In step 3 and 4, for each group at leaf node with multiple guideline-concordant medication options, we perform the drug effectiveness analysis and subgroup analysis on the data samples, and the results can be associated as the evidence. These evidence of the leaf node can be applied to prioritize the corresponding candidate medications, or once confirmed by domain experts, to refine the clinical guideline, i.e., to further split the leaf node into a sub-tree.

1. Applying medical data into the knowledge model and evaluating

Applying medical data into the knowledge model, we obtained seven groups. A group corresponds to a leaf node and contains data samples with similar features and different medication patterns in their prescriptions. Comparing the medication in physician’s prescription with standardized medications from knowledge model, data samples in each group can be divided into two kinds: “guideline-concordant” or “guideline-not-concordant”. Data samples were categorized to the guideline-not-concordant cohort if their prescription is not in any of standardized medications of this group. The standardization degree for medication and its relationship with effectiveness were evaluated by guideline adherence of prescriptions and the rate of achieving glycemic control goal.

![Figure 1. Methodology of integrating clinical knowledge and RWE for T2DM treatment.](image)
To reveal the relationship between guideline-concordant medication and its effectiveness indicated by glycemic control goal achieving rate, multivariate logistic regression model was applied to adjust for confounding factors. Given that the glycemic control goal achieving rate could be affected by the baseline blood glucose level, the baseline FPG was identified as a confounding factor. Adjusted odds ratios (ORs) of achieving glycemic control goal could be obtained via multivariate logistic regression. The same analysis was applied for each group.

2. Medication effectiveness analysis and subgroup analysis

Our approach enables the guideline adherence and personalization recommendation. In each group, we first compared the effectiveness of all guideline-concordant medication options at the whole data samples of the group, where the guideline-not-concordant medications were excluded. Similarly, adjusted OR, P-value in multivariate logistic regression model were used to reveal the relationship between medication options and glycemic control goal achievement. Here, we identified confounding factors by selecting the features that were associated with the medication options, from the basic information, physical measurement and history of diseases.

If the result of the medication comparative effectiveness analysis was not statistically significant at the group level, we would apply Model-based recursive partitioning approach (MOB) on the leaf node to further “grow” the tree for subgroup analysis.

The basic idea of MOB is that each node is associated with a single model. The datasets are split into different subsets based on partitioning variables to make the model parameters stable and the distributions of the response values are most different. That means a precision cohort fitted well or with strong evidence may be found after further partitioning and the drug effectiveness result of the subgroup determines the final recommended prescriptions.

The based parametric model in our MOB tree is logistic regression in which medication option (pairwise or multiple options) is as independent variable and glycemic control goal is as dependent variable. For instance, binary medication options: MET or SU/GLN, glycemic control goal: FPG<7.0mmol/L.

The iterative steps of MOB method are as below:

1) fitted the logistic regression with glycemic control goal as response variable and binary medication options as independent variable on all observations. Given n observations \( Y_i (i = 1, 2, ..., n) \), the model can be fitted by minimizing some objective function \( \Psi(Y, \theta) \) yielding the parameter estimate \( \hat{\theta} \). Maximum likelihood method or least squares method can obtain the solution of \( \hat{\theta} \).

\[
\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \Psi(Y_i, \theta), \quad \text{where} \quad Y_i = \begin{cases} 1, & \text{FPG} < 7.0\text{mmol/L} \\ 0, & \text{FPG} \geq 7.0\text{mmol/L} \end{cases}
\]

2) test the parameter instability on partitioning variables (age, gender, historical diseases and other demographic or medical characteristics).

3) if significant parameter instability (i.e., the smallest P values) is detected, split the sample with respect to the partitioning variable into two subsets.

4) repeat step (1) ~ (3) until there is no significant parameter instability on each node.

MOB is a recursive algorithm, since each subgroup will be further split until recursive end condition is reached.

Results

Data Set

We demonstrate the effectiveness of our method using a dataset that is a collection of type 2 diabetes patient records across multiple hospitals in a city of China with time spanning from 2015 to 2017. In total, there were 119236 records with the main diagnosis as type 2 diabetes, corresponding to 50061 unique patients. Each sample in the medical data recorded one visit of T2DM patient, which consisted of basic information of the patient, diagnosis made by the physician, blood glucose level (FPG), and the medication prescription. The distribution of age centered around 60 to 80 years old. Females accounted for 61% of the whole data. Patients with BMI more than 24 accounted for 42%. Those demographic characteristics described from data agree well with the fact that obese people are more susceptible to type 2 diabetes.

Medication Pattern Mining

Here we used association rule mining method (Aprior algorithms) to identify the frequent sets of medications with a support threshold of 0.01%. As shown in Table 2, top 10 medication patterns identified for type 2 diabetes treatment
were all consistent with the instruction from the knowledge model. Among all prescriptions, the most popular one was MET only, which accounted for about 46% of all samples. This is in consistent with the fact that MET is the first priority in T2DM therapy except for special cases. In addition, a combination of MET and SU/GLN was very common in Dual Therapy, which accounted for about 15% of all samples.

Table 2. Description of top 10 medication patterns for T2DM

<table>
<thead>
<tr>
<th>Medication Patterns*</th>
<th>Med_num</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MET</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>SU/GLN</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>SU/GLN+MET</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>MET+basal_insulin</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>SU/GLN+basal_insulin</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>AGIs</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>MET+AGIs</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>SU/GLN+MET+basal_insulin</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>SU/GLN+MET+AGIs</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>SU/GLN+AGIs</td>
<td>2</td>
</tr>
</tbody>
</table>

* In combination prescriptions, each drug category is connected using the “+” sign. Medication patterns showed in guideline are marked in bold.

Table 3. Statistical results of guideline adherence in medications for each group

<table>
<thead>
<tr>
<th>GroupID</th>
<th>Condition</th>
<th>Total number</th>
<th>Guideline-concordant number</th>
<th>Guideline-concordant ratio</th>
<th>Main guideline-not-concordant medication*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lifestyle Management Untreated, FPG&lt;7.0mmol/L</td>
<td>28107</td>
<td>14744</td>
<td>52.46%</td>
<td>Single drug(31.83%) Double drug(14.95%)</td>
</tr>
<tr>
<td>2</td>
<td>Monotherapy Untreated, 7.0mmol/L&lt;=FPG&lt;11.1mmol/L</td>
<td>10406</td>
<td>3478</td>
<td>33.42%</td>
<td>No drug(44.01%) Double drug(19.48%)</td>
</tr>
<tr>
<td>3</td>
<td>Short-term Intensive insulin Therapy Untreated, FPG&gt;=11.1mmol/L</td>
<td>1895</td>
<td>271</td>
<td>14.3%</td>
<td>Single drug(38.05%) No drug(37.47%)</td>
</tr>
<tr>
<td>4</td>
<td>Continued Therapy Treated, FPG&lt;7.0mmol/L</td>
<td>58195</td>
<td>56934</td>
<td>97.83%</td>
<td>Single drug(0.97%) Double drug(0.55%)</td>
</tr>
<tr>
<td>5</td>
<td>Dual Therapy Treated, FPG&gt;=7.0mmol/L, used single drug but no insulin</td>
<td>11767</td>
<td>191</td>
<td>1.62%</td>
<td>Single drug(97.50%) Double drug(0.76%)</td>
</tr>
<tr>
<td>6</td>
<td>Triple Therapy Treated, FPG&gt;=7.0mmol/L, used double drug but no insulin</td>
<td>4567</td>
<td>16</td>
<td>0.35%</td>
<td>Single drug(7.90%) Double drug(91.48%)</td>
</tr>
<tr>
<td>7</td>
<td>Combination Injectable Therapy Treated, FPG&gt;=7.0mmol/L, used three drugs or used insulin</td>
<td>4299</td>
<td>166</td>
<td>3.86%</td>
<td>Single drug(14.31%) Double drug(62.94%)</td>
</tr>
<tr>
<td>8</td>
<td>All</td>
<td>119236</td>
<td>75800</td>
<td>63.57%</td>
<td></td>
</tr>
</tbody>
</table>

* In guideline-not-concordant medication, only top 2 are showed. Items in bold are guideline-not-concordant ratios greater than 50%. 

842
Applying medical data into the knowledge model and evaluating result

Applying medical data into the knowledge model, we obtained seven groups. As shown in Table 3, by comparing the medication in physician’s prescription with standardized medications from knowledge model, the standardization degree for medication in each group was described by guideline-concordant number, guideline-concordant ratio and main guideline-not-concordant medication.

In summary, the majority (63.57%) of medications were concordant with the guideline. Especialy in “Continued Therapy” group whose instruction is maintaining previous treatment, the guideline-concordant ratio was up to 97.83% because prescriptions always follow previous prescriptions in the data. However, in “Dual Therapy” group, the guideline-concordant ratio was as low as 1.62% since 97.50% patient visits supposed to add a new drug category according to guideline were still prescribed with previous single drug category.

There were 67493(57%) records with FPG after about 3 months’ therapy. Based on these samples we evaluated the relationship between guideline-concordant medication and glycemic control goal achieving rate, as shown in Table 4.

Table 4. Association between guideline-concordant medication and blood glucose control for each group

<table>
<thead>
<tr>
<th>GroupID</th>
<th>Guideline-concordant</th>
<th>Total number</th>
<th>Number of not achieving FPG goal</th>
<th>Rate of not achieving FPG goal</th>
<th>Number of achieving FPG goal</th>
<th>Rate of achieving FPG goal</th>
<th>Base FPG mean</th>
<th>P value*</th>
<th>OR*</th>
<th>Note*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lifestyle Management yes</td>
<td>7787</td>
<td>1481</td>
<td>19.02%</td>
<td>6306</td>
<td>80.98%</td>
<td>6.02</td>
<td>&lt;0.001</td>
<td>1.20</td>
<td></td>
<td>significant (achieving) concordant &gt; not concordant</td>
</tr>
<tr>
<td>no</td>
<td>4402</td>
<td>967</td>
<td>21.97%</td>
<td>3435</td>
<td>78.03%</td>
<td>5.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monotherapy yes</td>
<td>1727</td>
<td>638</td>
<td>36.94%</td>
<td>1089</td>
<td>63.06%</td>
<td>8.28</td>
<td>&lt;0.001</td>
<td>2.48</td>
<td></td>
<td>significant (achieving) concordant &gt; not concordant</td>
</tr>
<tr>
<td>no</td>
<td>3670</td>
<td>2174</td>
<td>59.24%</td>
<td>1496</td>
<td>40.76%</td>
<td>8.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term Intensive insulin Therapy yes</td>
<td>148</td>
<td>124</td>
<td>83.78%</td>
<td>24</td>
<td>16.22%</td>
<td>14.23</td>
<td></td>
<td>0.65</td>
<td>0.50</td>
<td>no significant difference</td>
</tr>
<tr>
<td>no</td>
<td>795</td>
<td>572</td>
<td>71.95%</td>
<td>223</td>
<td>28.05%</td>
<td>13.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continued Therapy yes</td>
<td>36241</td>
<td>6629</td>
<td>18.29%</td>
<td>29612</td>
<td>81.71%</td>
<td>6.01</td>
<td>0.30</td>
<td>1.10</td>
<td></td>
<td>no significant difference</td>
</tr>
<tr>
<td>no</td>
<td>733</td>
<td>147</td>
<td>20.05%</td>
<td>586</td>
<td>79.95%</td>
<td>6.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dual Therapy yes</td>
<td>105</td>
<td>63</td>
<td>60.00%</td>
<td>42</td>
<td>40.00%</td>
<td>9.67</td>
<td>0.34</td>
<td>0.82</td>
<td></td>
<td>no significant difference</td>
</tr>
<tr>
<td>no</td>
<td>7052</td>
<td>3557</td>
<td>50.44%</td>
<td>3495</td>
<td>49.56%</td>
<td>8.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triple Therapy yes</td>
<td>11</td>
<td>5</td>
<td>45.45%</td>
<td>6</td>
<td>54.55%</td>
<td>9.15</td>
<td>0.60</td>
<td>1.38</td>
<td></td>
<td>no significant difference</td>
</tr>
<tr>
<td>no</td>
<td>3055</td>
<td>1615</td>
<td>52.86%</td>
<td>1440</td>
<td>47.14%</td>
<td>8.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combination Injectable Therapy yes</td>
<td>58</td>
<td>36</td>
<td>62.07%</td>
<td>22</td>
<td>37.93%</td>
<td>9.19</td>
<td></td>
<td>0.45</td>
<td>0.79</td>
<td>no significant difference</td>
</tr>
<tr>
<td>no</td>
<td>1709</td>
<td>970</td>
<td>56.76%</td>
<td>739</td>
<td>43.24%</td>
<td>9.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All yes</td>
<td>46077</td>
<td>8976</td>
<td>19.48%</td>
<td>37101</td>
<td>80.52%</td>
<td>6.17</td>
<td>&lt;0.001</td>
<td>3.62</td>
<td></td>
<td>significant (achieving) concordant &gt; not concordant</td>
</tr>
<tr>
<td>no</td>
<td>21416</td>
<td>10002</td>
<td>46.70%</td>
<td>11414</td>
<td>53.30%</td>
<td>7.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Items in bold are statistically significant at P<0.05 and OR>1.
According to the evaluation result, we found that guideline-concordant treatments was associated with better clinical outcome in FPG control than guideline-not-concordant ones. The contrast of FPG-goal achieving rates show that patients with guideline-concordant prescriptions acquire higher FPG-goal achieving rate next visit time. For total samples, the FPG-goal achieving rate for guideline-concordant prescriptions was as high as 80.52%, while the FPG-goal achieving rate for guideline-not-concordant prescriptions was just 53.30%, with P value <0.001 (SS) and a rather high adjusted OR=3.62.

As mentioned before, to reveal the relationship between guideline-concordant medication and glycemic control goal achieving rate, multivariate logistic regression model was applied to adjust for confounding factors. Here base FPG was regarded as a confounding factor, which stands for the baseline level of blood glucose and has effect on FPG goal achieving rate. The mean value of base FPG was also shown in Table 4. For example, in “Dual Therapy” group, by descriptive statistic a contrary finding was that patients with guideline-concordant prescriptions acquire lower FPG-goal achieving rate (40% vs 50%). By adjusting of confounding factor, such as the base FPG (mean base FPG for guideline-concordant: 9.67mmol/L vs guideline-not-concordant: 8.70mmol/L), the adjusted P-value shows the no significant difference between guideline-concordant medication and glycemic control goal achieving rate.

For the “Triple Therapy” and “Combination Injectable Therapy” groups, the sample size is greatly unbalanced, with much fewer samples consistent with the guidelines than those inconsistent with the guidelines. The results in these groups were considered with no significant difference.

**Medication effectiveness analysis and subgroup analysis result**

Our approach enables the guideline adherence and personalization recommendation. First, the clinical guideline often suggests several candidate drug classes/drug-class combinations for similar patients with certain clinical conditions. Then the real world evidence would serve as the information for prioritizing multiple candidates and even refining the guideline to finer subgroups via drug effectiveness analysis. It helps physicians make informed decisions. In each group, we need to compare the effectiveness of all guideline-concordant medication options based on the data samples of the group.

Taking the group of “Monotherapy” for example, the usage number of each guideline-concordant medication and its effectiveness (FPG goal achieving rate) was shown in Table 5 with descriptive statistic. Since the number of using AGIs is quite small, we give the comparative effectiveness analysis just between MET and SU/GLN. The result shows that the adjusted P value=0.34 and OR=0.895, which means there was no significant difference in effectiveness between MET and SU/GLN. Here BMI, age and history of hypertension were identified as the confounding factors to adjust OR, because they were associated with medication options or glycemic control goal achievement significantly.

**Table 5.** Descriptive statistic result of medication usage and effectiveness in “Monotherapy” group

<table>
<thead>
<tr>
<th></th>
<th>Total number</th>
<th>Number of not achieving FPG goal</th>
<th>Rate of not achieving FPG goal</th>
<th>Number of achieving FPG goal</th>
<th>Rate of achieving FPG goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>MET</td>
<td>1200</td>
<td>436</td>
<td>36.33%</td>
<td>764</td>
<td>63.67%</td>
</tr>
<tr>
<td>SU/GLN</td>
<td>498</td>
<td>191</td>
<td>38.35%</td>
<td>307</td>
<td>61.65%</td>
</tr>
<tr>
<td>AGIs</td>
<td>29</td>
<td>11</td>
<td>37.93%</td>
<td>18</td>
<td>62.07%</td>
</tr>
<tr>
<td>All</td>
<td>1727</td>
<td>638</td>
<td>36.94%</td>
<td>1089</td>
<td>63.06%</td>
</tr>
</tbody>
</table>

If the result of the comparative effectiveness analysis for medication options is not statistically significant at the group level, we apply recursive partitioning approach on the leaf node to further “grow” the tree for subgroup analysis. Here we grow a MOB tree on “Monotherapy” group to perform subgroup analysis, as shown in Figure 2. We set stopping condition: significant parameter instability P value<0.5, maximum depth of MOB tree maxDepth=3. The partitioning variables are detected after parameter instability test, they are BMI (kg/m²) and age (year). The partitioning point for BMI is 23.34 kg/m², 58 years for age variable. As shown in Figure 2A, Monotherapy group were split into three subgroups (cluster) according to BMI and age. Three single logistic regression model fitted on each cluster, and regression coefficients and adjusted ORs of treatment options are shown in Table 6. Finally, we compared the FPG goal achieving rate using MET or SU/GLN respectively, as shown in Figure 2B. Although it is not significant due to the small number of patient samples, SU/GLN can lead to better outcome in cluster 1. In cluster 2 and 3, MET can lead to better outcome which is consistent with the guideline (Improper use of SU may cause hypoglycemia especially
in elderly patients and also cause weight gain). According to the results of subgroup analysis, patients in cluster 1 (BMI≤23.34 kg/m², age≤58 years) are recommended to SU/GLN; patients in cluster 2 (BMI≤23.34 kg/m², age>58 years) are recommended to MET; for patients in cluster 3, recommendations are with no preference currently.

**Figure 2.** Result of subgroup analysis with recursive partitioning approach in Monotherapy group. (A) Sub tree for subgroup splitting, (B) Glucose control rate for each medication option in all group and each subgroup.

**Table 6.** Result of comparative effectiveness analysis on three subgroups (clusters) of Monotherapy group

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Adjust.OR</th>
<th>Coef</th>
<th>P value</th>
<th>Count</th>
<th>FPG goal achieving rate using MET</th>
<th>FPG goal achieving rate using SU/GLN</th>
<th>Precision recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster_1</td>
<td>1.64</td>
<td>0.50</td>
<td>0.24</td>
<td>124</td>
<td>58%</td>
<td>71%</td>
<td>SU/GLN</td>
</tr>
<tr>
<td>cluster_2</td>
<td>0.84</td>
<td>-0.17</td>
<td>0.36</td>
<td>704</td>
<td>72%</td>
<td>68%</td>
<td>MET</td>
</tr>
<tr>
<td>cluster_3</td>
<td>0.96</td>
<td>-0.04</td>
<td>0.79</td>
<td>812</td>
<td>57%</td>
<td>56%</td>
<td>MET or SU/GLN</td>
</tr>
<tr>
<td>All_group</td>
<td>0.95</td>
<td>0.53</td>
<td>-0.06</td>
<td>1664</td>
<td>63%</td>
<td>62%</td>
<td>/</td>
</tr>
</tbody>
</table>

**RWE-integrated CDSS**

A CDSS for type 2 diabetes treatment is built with real-world evidence. After collection of patient data including basic information, lab or examine results, previous treatment, historical diseases like comorbidity or complication, the CDSS system can recommend the medications and provide the evidence of why the recommendation is made, according to the knowledge model and data-based evidence of similar patients. As shown in the top area of Figure 3, the system provided comprehensive medications for T2DM patients including blood pressure management, blood glucose management, blood lipid management and blood platelet management respectively. And each kind of medication recommendation includes class name, generic name and product name. For patients with certain conditions, there could be more than one treatment options in every management module. Physicians can switch the treatment by clicking the drop-down menu. After clicking the evidence icon next to the medication options, both the data-based evidence and knowledge-based evidence are showed at the bottom. In the bottom-right area of Figure 3, for the data-based evidence in “similar patients” tab, feature statistic, using percentage and glucose control rate of medication options in similar patients are showed in the table, the pie chart and the bar chart respectively. As shown in the bottom-left area of Figure 3, the path in the decision tree is highlighted, indicating why the recommendation is made for this individual patient according to the clinical guideline. In the real application, the physicians can turn to the CDSS for 1) reasonable medication options according to our knowledge model and 2) more detailed evidence from similar patients’ data to prioritize the corresponding candidates. After choosing the medication treatment, the physician could further customize the treatment. He or she could select/unselect the drug class by the check box, and choose the drug with specific generic name and product name. Till now, we have performed the pilot testing on the CDSS system in several hospitals in China.

**Discussion**

In this study, we proposed a method and system by integrating clinical guideline and RWE for T2DM treatment. The clinical guideline provides standardized medications and RWE is generated to prioritize the corresponding candidates.
such as effectiveness in similar patients. For similar-patient group that is not statistically significant in effectiveness, we apply recursive partitioning approach to further “grow” the tree for subgroup analysis.

The strengths of this study include the integration of the clinical knowledge and real-world evidence, the achieving the standardized (i.e., consistent with knowledge) and personalized (i.e., referring to the RWE of similar patients) medication recommendation for T2DM patients, the medication effectiveness analysis and subgroup analysis, and finally the CDSS system realization and application.

The limitation of the current work is that evidence was based on drug categories in treatment. In fact, a physician need to decide which generic name or product name of drug can be most appropriate. Moreover, we only evaluated short term glycemic control goal achievement rather than long term outcome improvement such as the occurrence of complications. When more data are accumulated with longer time span, we can refine the recommendation and evidence to generic or product level and evaluate long term outcome improvement of the patients, so that the system will be more helpful and practical for physicians. Another limitation is that the outcome used to evaluate the glycemic control has been discretized as goal-achieving and goal-not-achieving, ordinal regression model for continuous outcome can be applied with more comprehensive evaluation.

Our approach is tested with type 2 diabetes patients’ data. The method can be generalized to other chronic diseases based on its guideline. By customization of medication pattern, features, effectiveness, and related parameters, our approach can be applied in different scenarios.

Figure 3. Screen shot of the RWE-integrated CDSS
Conclusion

In this paper, we proposed a method and developed a CDSS by integrating clinical knowledge and real-world evidence for T2DM treatment. We built our knowledge model based on Chinese clinical guideline for prevention and treatment of type 2 diabetes (2017) and integrated it with RWE extracted from more than 110,000 records of T2DM patients in a Chinese city. The knowledge model has been verified by the clinical experts from the advanced hospitals. The data verification results show that the medications that are consistent with the method recommendation can lead to better clinical outcome in terms of glycemic control, compared to those are inconsistent. Furthermore, the method and system enabled the personalized medication recommendation by comparing the effectiveness of treatment options in terms of the clinical outcomes like glycemic control goal achieving rate.

Reference

Assessing Contribution of Higher Order Clinical Risk Factors to Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage Patients

Azade Tabaie, MSc1, Shamim Nemati, PhD1, Jason W. Allen, MD, PhD2, Charlotte Chung, MD2, Flavia Queiroga, BSc2, Won-Jun Kuk, BSc3, Adam B. Prater, MD2
1Department of Biomedical Informatics, Emory School of Medicine, Atlanta, GA, USA; 2Department of Radiology, Emory School of Medicine, Atlanta, GA, US; 3Emory School of Medicine, Atlanta, GA, US;

Abstract

The goal of this study was to investigate the application of machine learning models capable of capturing multiplicative and temporal clinical risk factors for outcome prediction in patients with aneurysmal subarachnoid hemorrhage (aSAH). We examined a cohort of 575 aSAH patients from Emory Healthcare, identified via digital subtraction angiography. The outcome measure was the modified Ranking Scale (mRS) after 90 days. Predictions were performed with longitudinal clinical and imaging risk factors as inputs into a regularized Logistic Regression, a feedforward Neural Network and a multivariate time-series prediction model known as the long short-term memory (LSTM) architecture. Through extraction of higher-order risk factors, the LSTM model achieved an AUC of 0.89 eight days into hospitalization, outperforming other techniques. Our preliminary findings indicate the proposed model has the potential to aid treatment decisions and effective imaging resource utilization in high-risk patients by providing actionable predictions prior to the development of neurological deterioration.

Introduction and Background

Aneurysmal subarachnoid hemorrhage (aSAH) is a grave medical condition, affecting over 30,000 individuals a year in the United States1. Despite advances in medical care, it remains a major cause of premature mortality, accounting for 27% of all stroke-related years of life lost before the age of 652. Mortality is estimated at 30%, and up to 40% of survivors have long-term neurological deficits resulting in significant loss of quality of life and increased burden and cost on the healthcare system3. Delayed cerebral ischemia (DCI) and cerebral infarction are major complications of aSAH, occurring in 19-46% of patients, and account for a large burden of aSAH-related morbidity4,5. Cerebral arterial vasospasm (CVS) is another major complication of aSAH, seen in up to 70% of patients, and is often associated with DCI and cerebral infarction6. For this study, CVS is defined as arterial narrowing documented on imaging, specifically computed tomography angiography (CTA) and/or digital subtraction angiography (DSA). As DCI is often difficult to diagnose and predict, the strong association between the DCI and CVS has led to the clinical practice of using CVS detection as a means of early detection of DCI, in an effort to guide early treatment to prevent cerebral infarction. However, less than half of patients with CVS ultimately develop DCI, thus definitive prognostic and therapeutic decision based solely on CVS screening is not ideal and may lead to overtreatment and treatment-related morbidity. Furthermore, while less common, DCI also can occur in the absence of CVS further complicating timely diagnosis and treatment7,8.

There are a variety of imaging studies used to diagnose CVS, each with its own distinct advantages and pitfalls, including transcranial doppler ultrasound (TCD), computed tomographic angiography (CTA) and CT perfusion (CTP). However, there is conflicting evidence in the literature regarding sensitivity, specificity, and prognostic efficacy. In particular, TCD has been deemed both effective8,10 and ineffective11,12 in predicting both vasospasm and poor neurologic outcomes, using a variety of imaging parameters and outcome definitions. CTA and CTP are additional imaging modalities used to diagnose vasospasm and define brain tissue at risk for DCI. These imaging modalities have been shown to have a higher sensitivity for detection of vasospasm and DCI and have generally been found useful for prognostication13,14. However, the radiation exposure from these techniques must be considered. DSA is considered the gold standard for determination of angiographic vasospasm, but it is an invasive and resource intensive modality with a small, but significant risk of neurologic compromise. A review of the literature to evaluate the diagnostic value of DSA for detection of DCI revealed few studies14. Furthermore, a recent study found that screening asymptomatic patients with DSA was ineffective, although this remains useful in patients that are either symptomatic or difficult to assess clinically due to poor neurologic status15.
A wide variety of clinical signs and risk factors for developing both CVS and DCI have been described, including smoking\textsuperscript{16, 17}, hypertension\textsuperscript{16, 17}, hyperglycemia\textsuperscript{16, 17}, old age\textsuperscript{18, 19}, and level of consciousness\textsuperscript{20, 21}, as well as image-based scoring systems such as modified Fisher score (MFS)\textsuperscript{22, 23}. However, much of the literature is difficult to interpret as outcomes are often poorly defined\textsuperscript{24}. In fact, the terms CVS and DCI are often used interchangeably, further complicating study comparison. Variations in study design and differing outcome definitions underscore the need for a robust, reproducible predictive model with clearly defined outcomes. Additionally, the complex pathophysiology of aSAH, diagnostic challenges of DCI, and risk of poor neurologic outcome emphasizes the need for a prognostic tool to aid in the early identification of aSAH patients at risk for vasospasm, ischemia and stroke prior to neurologic deterioration.

In clinical research, conventional statistical techniques such as linear and logistic regression are the dominant techniques used for clinical decision support models. These techniques are hypothesis driven and simple to interpret, yet overly simplistic statistical assumptions about data distribution and model fitting, as well as inability to identify higher order interactions among the input variables, limit their applicability in complex data sets. Notably, logistic regression has been used to evaluate the contribution of TCD to predicting outcomes in aSAH patients, and the literature reveals conflicting results, with TCD shown to be both contributory\textsuperscript{11} and noncontributory\textsuperscript{11, 25}. These findings reinforce the need for an alternative approach to data modeling in aSAH patients.

Machine learning is a branch of artificial intelligence that has been employed in a variety of applications, including medicine, to analyze complex datasets\textsuperscript{26}. While the complexity of clinical problems such as aSAH outcome prediction provides an opportunity for machine learning techniques, the smaller datasets available in healthcare compared to other industries provide a significant challenge to machine learning techniques\textsuperscript{27–29}.

There have been a few preliminary studies of machine learning models to predict vasospasm in patients with aSAH, which outperformed their respective logistic regression models\textsuperscript{30–32}. However, two of the models were trained on less than 100 patients, limiting generalizability. Additionally, focusing on vasospasm as an outcome, rather than a measurable clinical outcome, limited the clinical utility of these models. The third model, while trained on a large dataset from multiple sites, did not include several important variables previously identified as prognostic for poor outcome, such as smoking status\textsuperscript{32}.

The superior performance of artificial neural networks (ANN) on complex datasets make them well-suited for clinical applications including longitudinal data and multiplicative risk factors. A Recurrent neural network (RNN) is a special type of ANN tailored towards finding predictive patterns in longitudinal multivariate data, due to their use of internal states (or memory) to process sequences of inputs. This internal memory, and the ability to model trajectories, make RNNs a powerful tool for monitoring of patients staying in hospitals to prevent adverse events\textsuperscript{33}.

In this work, we compare three predictive modeling approaches, namely a logistic regression model, an ANN model, and an RNN model, to test the hypothesis that multiplicative risk factors (or interaction terms), and temporal features in the data can improve prediction of outcomes in aSAH patients. The rest of this paper is organized as follows. We start with a description of our aSAH patient cohort and the utilized machine learning models, results, and significance of our findings and future directions.

Data and Methods

This investigation was conducted according to Emory University Institutional Review Board approved protocol 66,389. Our preliminary dataset identified approximately 4,191 patients with confirmed aSAH by DSA dating back to January 1998. We used a subset containing 575 patients dating back to January 2009 for this preliminary study. The Emory Healthcare Clinical Data Warehouse (CDW) was queried for the 575 patients in our data subset, with data extraction query for demographic, clinical and imaging data identified in our literature review as associated with aSAH outcome in prior prognostic models. Associated clinical imaging studies and reports included in the CDW retrieval are: TCD, non-contrast head CT head CTA, head CTP, and DSA. Two board-certified neuroradiologists also reviewed each patients presenting head CT to determine MFS. Individual chart-review supplemented clinical factors not expected to be captured in the CDW retrieval, including level of consciousness, and World Federation of Neurosurgeons grade. The specific model features extracted were: hypertension\textsuperscript{16, 17}, smoking status\textsuperscript{16}, hyperglycemia\textsuperscript{16}, diabetes mellitus\textsuperscript{16}, increased intracranial pressure\textsuperscript{19}, leukocytosis\textsuperscript{34}, level of consciousness\textsuperscript{20, 21}, weight\textsuperscript{17}, World Federation of Neurosur-
geons grade\textsuperscript{35,36}, and blood transfusion\textsuperscript{37}, MFS\textsuperscript{22,23}, aneurysm treatment (surgical clipping or endovascular coil)\textsuperscript{38}, presence of perfusion mismatch on CTP\textsuperscript{13,14,39}, TCD vessel elevation $> 140 cm/sec$\textsuperscript{11}, CVS on CTA\textsuperscript{14,15}, CVS on DSA\textsuperscript{14,15}, aneurysm location, number of aneurysms and aneurysm size\textsuperscript{37}. Inclusion criteria are patients 18 years of age or older with documented aSAH at admission based on the results of head CTA or DSA. Exclusion criteria are causes of SAH other than aneurysm including but not limited to trauma, infection, vascular malformation, vasculitis, post surgical complication, reversible cerebral vasospasmodic syndrome, and Moya Moya.

Two general types of variables are included in the model; time-variant features which includes laboratory values, vital signs and imaging exams, and time-invariant features which contains the information regarding demographic and patient history. Some variables such as vital signs and laboratory values are updated more frequently comparing to imaging exams. Each patient information was binned with 24-hour time intervals. Sample-and-hold technique was applied on the data. If there were more than one value for imaging exams, we took the median and if there were more than one value for vital signs and laboratory results, we took the mean of the available values. Missing values were imputed with the mean or median of the corresponding feature. Then, numerical features were normalized and transformed if needed.

The outcome labels for this study were the modified ranking scale (mRS) which is a commonly used scale for measuring the degree of disability or dependence in the daily activities of people who have suffered a stroke or other causes of neurological disability. The mRS score ranges from zero to six, where an mRS of zero corresponds to having no symptoms of disability while an mRS score of six indicates that the patient has expired. We dichotomized the mRS scores into binary outcomes\textsuperscript{24}, corresponding to poor (mRS = 4, 5, 6) and favorable (mRS = 0, 1, 2, 3) outcomes. This measure was used as a surrogate of DCI and cerebral infarction, as this measure is a common secondary outcome used as a surrogate for DCI\textsuperscript{40}.

In order to predict patients mRS each day, the data was binned based on 24 hour time intervals. Therefore, for each patient we had a sequence of daily information starting from the end of the first day of ICU admission. In this cohort, each patient had only one outcome which was assigned 90 days after being discharged from the hospital. In order to make daily prediction for each patient, the outcome was replicated\textsuperscript{41} and assigned to every day that patient was hospitalized in the ICU.

Our initial model consisted of logistic regression with L1 and L2 regularization (also known as an Elastic Net model) with 10-fold internal cross-validation to optimize the penalty term. We hypothesized that both a neural network model and a recurrent neural network model with long short-term memory (LSTM) would provide more robust prediction in patients with aSAH when compared to logistic regression.

Traditional ANNs were not designed to analyze sequential data, which complicates analysis of longitudinal data sets in clinical settings. An RNN model can be thought of as multiple copies of the same network, each passing a message to a successor. This chain-like structure and the associated memory make the RNN models ideal for use with longitudinal data. To make daily predictions of the outcome of aSAH patients, we utilized a Long Short-Term Memory (LSTM) network, which is capable of learning long-term dependencies in time series data\textsuperscript{42}.

The RNN model includes input layer containing sequential data, a LSTM layer with 50 hidden units, a fully connected layer followed by a Softmax layer and a classification layer. Number of hidden units was experimentally assigned, and a fully connected layer is added to capture all the interactions and nonlinear relationships among the features. To train the RNN model, mini batch technique was incorporated with mini batch size of 250, almost the maximum value possible. Adam optimizer was applied and the rest of the parameters in training process are initial learning rate=0.01, Gradient threshold=1, and Learning rate drop period=20. Moreover, a validation set was assigned to apply L2 regularization and prevent overfitting.

In order to make a fair comparison between the RNN model and the artificial neural network, we kept the same setting to train ANN, and only the LSTM layer was excluded from the model.

**Results**

Three predictive models were applied, namely an Elastic Net model, a feedforward ANN model with a fully connected layer and an LSTM-type recurrent neural network. To present the predictive power of LSTM, we compared the results
Figure 1: (a) AUC-ROC of applying the feedforward Artificial Neural Network on the training and testing sets (10-fold cross-validated). All patient time series were aligned at admission and risk scores were calculated on a daily basis for up to 15 days. (b) Performance of the L1L2 regularized Logistic Regression model (aka, Elastic Net).

The greatest risk of morbidity and mortality occurs during days 2-15 following initial aSAH hemorrhage. Therefore our model prediction focused on days 2-15 for outcome prediction. The performance measure we selected was area under the receiver operating curve (AUC-ROC or AUC) which is useful for comparing binary outcome classification models. The result of applying the trained models on different days of patient stay to predict the associated mRS is presented in Table 1 and Figure 2. Figure 2 demonstrates the average AUC of applying 10-fold cross-validation while predicting mRS on each day of patient stay in the ICU using the LSTM model. The vertical dotted lines present the error bar which is reasonably bounded. According to Figure 2, the overfitting is negligible and the LSTM model is performing well on both training and testing subsets. This model is capable of capturing non-linearity and temporal information in the data, and therefore in general we see an improvement in performance with inclusion of additional days of data. The best performance of the LSTM model is from day 8 to 13 in which AUC of 0.89 was achieved on testing set. Figure 1 (a) presents the average AUC from applying a NN model with 10-fold cross-validation to assess generalization performance of the model. The feedforward neural network models do not consider the temporal information in the data, but are capable of modeling non-linearity and interactions among the input features. The best performance of the NN model is on days 8, 11, and 12 which is 0.84. Figure 1 (b) shows the AUC of the ElasticNet model, which performed poorly compared to the NN and LSTM models. The main reason is that ElasticNet can neither capture temporal information nor non-linearity in the dataset. As a result, while this model is a regularized version of logistic regression model, ElasticNet is not capable of making meaningful predictions after day 5. In fact, the ElasticNet model severely overfits to the training data and performs poorly on the testing data, which is likely due to the smaller number of patient records available after day 5.

The details of predictive performance associated with the LSTM model, neural networks and ElasticNet model are presented in Table 1. The LSTM model constantly achieved better AUCs compared to the other models. Day 8 has the first best prediction performance for the LSTM model, so by testing the information on day 8 of the patient stay at ICU, we can obtain 0.89 prediction AUC. Time-dependent and time-independent features were employed in training the models. We performed sensitivity analysis in order to find the features which contributed most significantly to the AUC of the LSTM model. This analysis indicated that MRI and DSA features from imaging exams, INR, Phosphorus, white blood cells, Bilirubin, and PaO₂ from laboratory results, admission status, gender, and etiology category were the top features that contributed to the AUC achieved with the LSTM model.
Discussion and Concluding Remarks

Our results confirm our hypothesis that machine learning models can outperform traditional logistic regression in patients with aSAH. We believe that nonlinear interactions between variables, as well as the temporal relationships in longitudinal data are important to incorporate into a prediction model for this complex patient population. Our machine learning model design incorporates daily outcome prediction, which would theoretically allow physicians to more closely monitor patients at highest risk for poor outcome when used prospectively. Additionally, by focusing on an outcome related to DCI, our model does not solely rely on imaging features to potentially guide patient treatment decisions.

While this study was limited to a single institution and will require further validation, our results supports future exploration of the use of multivariate time-series prediction models such as LSTM neural networks on a larger, multi-institutional aSAH dataset. Further work could also explore the incorporation of live patient vital monitoring systems, which provide rich real-time datasets, as well as additional clinical outcomes such as mortality or cerebral infarction on follow up imaging. To eliminate the need for manual image annotation in a real-time implementation, one may build separate machine learning models to automatically annotate the images or use multimodal data fusion using a combination of convolutional neural networks for image feature extraction and recurrent neural networks for analysis of temporal information in time series data. The clinical implications of a successful predictive model for aSAH are profound, with the potential to reduce unnecessary radiation from inappropriate imaging, reduce hospital stay in low risk patients, and reduce morbidity in high risk patients. Further work in this area is warranted.

The results of this preliminary analysis demonstrate that a RNN with LTSM architecture provides robust evidence for clinical outcome prediction in patients with aSAH compared to traditional statistical techniques. We believe this supports further exploration with our full dataset, as well as expansion into a multicenter trial with a more diverse patient population.
Table 1: AUC-ROC of applying the trained LSTM, neural network and ElasticNet models to the training and testing sets. 10-fold cross-validation was applied, and the mean AUC across all folds is presented. Number of Samples denotes the available number of training and testing records as the number of days included in the analysis increases (note the drop in the number of records is due to death or discharge).

<table>
<thead>
<tr>
<th>Number of Days</th>
<th>Number of Samples</th>
<th>RNN (LSTM Model)</th>
<th>Neural Network</th>
<th>ElasticNet Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Testing</td>
<td>Training AUC</td>
<td>Testing AUC</td>
</tr>
<tr>
<td>1</td>
<td>459</td>
<td>116</td>
<td>0.84</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>458</td>
<td>115</td>
<td>0.88</td>
<td>0.82</td>
</tr>
<tr>
<td>3</td>
<td>456</td>
<td>115</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>4</td>
<td>453</td>
<td>114</td>
<td>0.91</td>
<td>0.86</td>
</tr>
<tr>
<td>5</td>
<td>447</td>
<td>112</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td>6</td>
<td>440</td>
<td>110</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td>7</td>
<td>434</td>
<td>109</td>
<td>0.93</td>
<td>0.88</td>
</tr>
<tr>
<td>8</td>
<td>427</td>
<td>107</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>9</td>
<td>414</td>
<td>104</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>400</td>
<td>100</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>11</td>
<td>376</td>
<td>94</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>12</td>
<td>354</td>
<td>89</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>13</td>
<td>321</td>
<td>81</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>14</td>
<td>298</td>
<td>75</td>
<td>0.92</td>
<td>0.88</td>
</tr>
<tr>
<td>15</td>
<td>275</td>
<td>70</td>
<td>0.92</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 2: p-values from pair-wise comparisons between the AUC-ROCs on the testing sets: 1) RNN versus ElasticNet, 2) NN versus ElasticNet. Three methods of comparing classifier performance were applied, namely Net Reclassification Improvement (NRI), Integrated Discrimination Improvement (IDI) and pauc\(^{13}\). p-values below 0.05 are considered significant, indicating statistically significant improvement in classification performance using RNN over ElasticNet, etc.

<table>
<thead>
<tr>
<th>Number of Days</th>
<th>RNN vs. ElasticNet</th>
<th>Neural Network vs. ElasticNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NRI</td>
<td>IDI</td>
</tr>
<tr>
<td>1</td>
<td>0.71</td>
<td>0.53</td>
</tr>
<tr>
<td>2</td>
<td>0.96</td>
<td>0.77</td>
</tr>
<tr>
<td>3</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>4</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>5</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>6</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>7</td>
<td>0.02</td>
<td>0.006</td>
</tr>
<tr>
<td>8</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>9</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>10</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>11</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>12</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td>13</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>15</td>
<td>0.14</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Acknowledgments
This study was supported by the Radiology Society of North America (RSNA) Seed Grant.
References


De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models

Buzhou Tang, PhD1*, Dehuan Jiang, MS1, Qingcai Chen, PhD1, Xiaolong Wang, PhD1, Jun Yan, PhD2, Ying Shen, PhD3

1Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology, Shenzhen, China
2Yidu Cloud (Beijing) Technology Co., Ltd, Beijing, China
3Ying Shen, Peking University, Shenzhen Graduate School, Shenzhen, China
*Corresponding author: tangbuzhou@gmail.com

Abstract

De-identification of clinical text, the prerequisite of electronic clinical data reuse, is a typical named entity recognition (NER) problem. A number of state-of-the-art deep learning methods for NER, such as Bi-LSTM-CRF (bidirectional long-short-term-memory conditional random fields), have been applied for de-identification. Neural language models used for language representation bring great improvement in lots of NLP tasks when they are integrated with other deep learning methods. In this paper, we introduce Bi-LSTM-CRF with neural language models for de-identification of clinical text, and evaluate it on the de-identification datasets of the i2b2 2014 and the CEGS N-GRID 2016 challenges. Four neural language models of three types individually integrated with Bi-LSTM-CRF are compared in this study. Bi-LSTM-CRF with neural language models achieves the highest “strict” micro-averaged F1-score of 95.50% on the i2b2 2014 dataset and 91.82% on the CEGS N-GRID 2016 dataset, becoming new benchmark results on these two datasets respectively.

Keywords: De-identification, Named entity recognition, Bidirectional long-short-term-memory, Conditional random fields, Neural language models.

Introduction

Clinical records as an important information source for medical research and investigations have been attracting more and more attention of medical professionals. A prerequisite for clinical records sharing is removing individually identifiable health information, known as protected health information (PHI). In the United States, the Health Insurance Portability and Accountability Act (HIPAA)[1] defines 18 different types of PHI. In clinical records, there is plenty of PHI embedded in clinical text, which cannot be identified directly like PHI recorded in structured table according to table fields.

The process of finding and removing PHI is called de-identification. As removing PHI is easy to implement, de-identification mainly focuses on finding PHI, and is usually recognized as a named entity recognition task. In the last decades, a variety of methods have been proposed for de-identification. These methods can be classified into two categories: rule-based methods and machine learning methods. Rule-based methods rely on manually constructed regular expressions and domain dictionaries, which are challenging and time-consuming to develop. They are easy to implement and do not require any labeled data. The main disadvantage of rule-based methods lies in that they are not easy to replant from one domain to another. To prevent disadvantages of rule-based methods, researchers have attempted machine learning methods for de-identification of clinical text, especially since several challenges about de-identification were organized, such as the i2b2 2014 challenge [2],[3] and the CEGS N-GRID 2016 challenge[4]. Classical statistical learning methods, such as decision trees [5], support vector machines [6], conditional random fields (CRF)[7] and structured support vector machines [8], have been applied for de-identification. Among them, CRF is one of the most state-of-the-art methods. All these methods require time-consuming feature engineering.

In recent years, deep neural networks, which have the ability of learning effective features from large-scale unlabeled data instead of feature engineering in traditional statistical learning methods, have been widely used in various tasks in natural language processing, such as language representation, NER, parsing, text classification, question answering and machine translation, and have shown promising results. For de-identification, deep neural networks also achieve state-of-the-art results [9],[10]. The representative method is Bi-LSTM-CRF (bidirectional long-short-term-memory conditional random fields)[10-15]. Bi-LSTM-CRF takes as input the sequence of word embeddings
and uses Bi-LSTM [16] for sentence representation and CRF for label sequence prediction. The input sequence of word embeddings learnt by neural language models on large-scale unlabeled data is an important influencing factor of Bi-LSTM-CRF. The neural language models may be classified into the following three types: 1) The common neural language models such as CBOW[17], Skip-gram[18] and GloVe[19] are usually used to obtain representations of words. Each word learnt by these models has only one representation, denoted by word embedding. However, a great percentage of words have different meanings when appearing in different context. 2) Recently, several neural language models that determine the embeddings of a word according to its context at each time dynamically, such as ELMo[20], GPT[21] and BERT[22], have been proposed and have brought significant improvement to existing methods. 3) Another way to improve existing methods using neural language models is optimizing neural language models and existing methods simultaneously. For example, Liu L et al.[23] introduced a character-level Bi-LSTM to predict each word of the input sentence, which was optimized together with Bi-LSTM-CRF. In this way, Bi-LSTM-CRF obtained improvement on a number of NLP tasks.

In this paper, we investigate Bi-LSTM-CRF with neural language models of the above-mentioned three types on de-identification and compare four neural language models, that is, Skip-gram[18], ELMo[20]. BERT[22] and LM[23]. Experiments conducted on the de-identification datasets of the i2b2 2014 challenge and the CEGS N-GRID 2016 challenge show that Bi-LSTM-CRF with neural language models achieves state-of-the-art performance. Bi-LSTM-CRF(BERT) is the best one with a “strict” micro-averaged F1-score of 95.50% on the i2b2 2014 de-identification dataset and 91.82% on the CEGS N-GRID 2016 challenge dataset, which is higher than that of the current best system[9] by 0.39%.

Methods

We start with the basic Bi-LSTM-CRF, and then the variants when different neural language models are added separately. Figure 1 shows their overall architecture.

Figure 1. Overview architecture of Bi-LSTM-CRF with neural language models

Bi-LSTM-CRF

It consists of three main components, called “layers” here, that are input layer, Bi-LSTM layer and CRF-layer. The input layer converts an input sentence $s=w_1w_2...w_n$ into a sequence of word embeddings $x=x_1x_2...x_n$. The Bi-LSTM layer further obtains two sequences $h_f=h_{f1}h_{f2}...h_{fn}$ and $h_b=h_{b1}h_{b2}...h_{bn}$, which represent context information of each
word of interest at every position from forward and backward directions respectively. The CRF layer finally predicts a label sequence $y=y_1y_2…y_n$. For detailed information, please refer to [13].

**Bi-LSTM-CRF with Neural Language Models**

There are two ways to integrate neural language models into Bi-LSTM-CRF: at the input layer and at the CRF layer. At the input layer, there are two types of representations for input sentences: word-level and sentence-level. At the CRF-layer, the input sentences should be also produced by neural networks like autoencoding.

Skip-gram is a model for word-level representations, the idea of which is to use a word of interest to predict words around it. Given large-scale unlabeled data, Skip-gram produces one embedding for each word. Then any input sentence can be represented by replacing each word by its word embedding, which is a process of dictionary lookup.

ELMo is a model for sentence-level representation. It uses stacked bidirectional LSTMs to encode input sentences, each word of which is predicted by the forward LSTM and the backward LSTM based on the two subsentences before and after it respectively. The same word in different sentences may have different meanings and should have different representations.

Similar to ELMo, BERT is also a model sentence-level representation. It adopts neural networks based on transformer to encode input sentences, where some of the words from input sentences are randomly masked, and are predicted based only on its context. Other than that, it further introduces a “next sentence prediction” task for text pair representation.

LM is a task-aware neural language model, which predicts each word based on character-level context and is optimized together with a label sequence task in a multi-task framework. When LM is integrated into Bi-LSTM-CRF at the CRF layer, another neural language model such as Skip-gram is also integrated at the input layer. In this study, we only consider the case of Bi-LSTM-CRF(Skip-gram+LM).

**Experimental Settings and Results**

**Datasets**

We compare different Bi-LSTM-CRF variants on the de-identification datasets of the i2b2 2014 challenge and the CEGS N-GRID 2016 challenge, which are publicly available. The i2b2 2014 dataset includes a training set of 790 records with 17,045 PHI instances and a test set of 514 records with 11,462 PHI instances. The CEGS N-GRID 2016 dataset includes of a training set of 790 records with 20,845 PHI instances and a test set of 400 records with 13,521 PHI instances. All PHI instances are classified into seven main categories with subcategories. The number of PHI instances in each category is listed in Table 1, where NA denotes no subcategory, and categories defined in HIPAA are marked by *.

**Experimental Settings**

**Evaluation:** The performance of all models is measured by standard micro-averaged precision (P), recall (R) and F1-score (F1) under different criteria, namely “token”, “strict”, “relaxed”, “HIPAA token”, “HIPAA strict”, “HIPAA relaxed”, “binary token”, “binary strict”, “binary HIPAA token” and “binary HIPAA strict”, where “token” and “strict” correspond to exact matching at token-level and instance-level respectively, “relaxed” allows at most two characters mismatched at the end, “HIPAA” corresponds HIPAA-defined categories, and “binary” only considers the boundaries of PHI instances no matter their categories. “strict” is the primary criterion. We use the same preprocessing as our previous work [9] in this study.

**Parameters:** We adopt the same settings as [9] for Bi-LSTM-CRF(Skip-gram) except that all categories of PHI instances are recognized by Bi-LSTM-CRF(Skip-gram) in this study, while some categories of PHI instances are recognized by rules in [9]. We use the pretrained ELMo (https://s3-us-west-2.amazonaws.com/AllenNLP/models/elmo/2x4096_512_2048cnn_2xhighway/elmo_2x4096_512_2048cnn_2xhighway_weights.hdf5) and BERT-Base (https://storage.googleapis.com/bert_models/2018_11_03/multilingual_L-12_H-768_A-12.zip) for neural language model initialization and they are fine-tuned during Bi-LSTM model training. The hyperparameters used in our study are listed in Table 2 in detail.
Table 1. Number of PHI instances in each category in the two de-identification datasets.

<table>
<thead>
<tr>
<th>Category</th>
<th>i2b2 2014</th>
<th>CEGS N-GRID 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
</tr>
<tr>
<td>NAME</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PATIENT*</td>
<td>1317</td>
<td>881</td>
</tr>
<tr>
<td>DOCTOR</td>
<td>2894</td>
<td>1913</td>
</tr>
<tr>
<td>USERNAME</td>
<td>264</td>
<td>92</td>
</tr>
<tr>
<td>PROFESSION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COUNTRY</td>
<td>66</td>
<td>117</td>
</tr>
<tr>
<td>STATE</td>
<td>312</td>
<td>190</td>
</tr>
<tr>
<td>CITY*</td>
<td>394</td>
<td>260</td>
</tr>
<tr>
<td>STREET*</td>
<td>216</td>
<td>136</td>
</tr>
<tr>
<td>ZIP*</td>
<td>212</td>
<td>140</td>
</tr>
<tr>
<td>LOCATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COUNTRY</td>
<td>442</td>
<td>524</td>
</tr>
<tr>
<td>STATE</td>
<td>611</td>
<td>422</td>
</tr>
<tr>
<td>CITY*</td>
<td>261</td>
<td>195</td>
</tr>
<tr>
<td>STREET*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ZIP*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDICALRECORD*</td>
<td>309</td>
<td>215</td>
</tr>
<tr>
<td>SSN*</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>DEVICE*</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>IDNUM*</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BIOID</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HEALTHPLAN*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VEHICLE*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ACCOUNT*</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LICENSE*</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Hyperparameters used in our study.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension of token-level word embeddings</td>
<td>Skip-gram: 50, ELMo: 1024, BERT: 768</td>
</tr>
<tr>
<td>Dimension of character-level word embeddings</td>
<td>Skip-gram: 25</td>
</tr>
<tr>
<td>Character-level LSTM size</td>
<td>Skip-gram: 25</td>
</tr>
<tr>
<td>Token-level LSTM size</td>
<td>100</td>
</tr>
<tr>
<td>Dropout probability</td>
<td>0.5</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.005</td>
</tr>
<tr>
<td>Training epochs</td>
<td>80</td>
</tr>
</tbody>
</table>

Experimental Results

The comparison results of the Bi-LSTM-CRF with different neural language models on the two datasets are shown in Table 3 and Table 4 respectively, where the highest P, R and F1 are highlighted in bold. The Bi-LSTM-CRF with the two neural language models, namely ELMo and BERT, for sentence-level representation achieves higher F1-score than the Bi-LSTM-CRF with the one neural language model, namely Skip-gram, for word-level representation. Bi-LSTM-CRF(BERT) achieves the highest “strict” F1-score of 95.50% on the i2b2 2014 dataset and 91.82% on the CEGS N-GRID 2016 dataset, outperforming LSTM-CRF(ELMo) by 1.54% and 1.12% respectively.
brings an improvement of 0.72% in “strict” F1-score on the i2b2 2014 dataset and 1.19% on the CEGS N-GRID 2016 dataset, when it is added into Bi-LSTM-CRF(Skip-gram). LSTM-CRF(ELMo) and Bi-LSTM-CRF(Skip-gram+LM) shows comparative performance.

Table 3. Comparison of Bi-LSTM-CRF variants with different neural language models on the i2b2 2014 dataset.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Bi-LSTM-CRF (Skip-gram)</th>
<th>Bi-LSTM-CRF (ELMo)</th>
<th>Bi-LSTM-CRF (BERT)</th>
<th>Bi-LSTM-CRF (Skip-gram+LM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P/R/F1(%)</td>
<td>P/R/F1(%)</td>
<td>P/R/F1(%)</td>
<td>P/R/F1(%)</td>
</tr>
<tr>
<td>Token</td>
<td>96.79/95.91/96.35</td>
<td>96.96/96.76/96.86</td>
<td>97.79/97.16/97.48</td>
<td>97.49/96.55/97.02</td>
</tr>
<tr>
<td>Strict</td>
<td>94.58/93.35/93.96</td>
<td>94.76/94.14/94.45</td>
<td>95.99/95.02/95.5</td>
<td>95.56/93.81/94.68</td>
</tr>
<tr>
<td>Relaxed</td>
<td>94.74/93.51/94.12</td>
<td>94.29/94.3/94.61</td>
<td>96.17/95.2/95.69</td>
<td>95.78/94.03/94.9</td>
</tr>
<tr>
<td>Binary token</td>
<td>98.81/97.91/98.36</td>
<td>98.57/98.37/98.47</td>
<td>99.02/98.38/98.7</td>
<td>99.06/98.1/98.58</td>
</tr>
<tr>
<td>Binary strict</td>
<td>96.63/95.38/96.0</td>
<td>96.59/95.96/96.28</td>
<td>97.35/96.36/96.85</td>
<td>97.28/95.51/96.39</td>
</tr>
<tr>
<td>HIPAA token</td>
<td>97.71/97.57/97.64</td>
<td>98.11/98.12/98.12</td>
<td>98.8/98.28/98.54</td>
<td>98.46/98.02/98.24</td>
</tr>
<tr>
<td>HIPAA strict</td>
<td>96.13/95.66/95.94</td>
<td>96.42/96.26/96.33</td>
<td>97.52/96.74/97.13</td>
<td>97.08/96.25/96.66</td>
</tr>
<tr>
<td>HIPAA relaxed</td>
<td>96.26/95.79/96.02</td>
<td>96.57/96.39/96.48</td>
<td>97.66/96.88/97.27</td>
<td>97.2/96.36/96.78</td>
</tr>
<tr>
<td>HIPAA binary token</td>
<td>98.08/97.94/98.01</td>
<td>98.35/98.36/98.36</td>
<td>98.98/98.45/98.71</td>
<td>98.68/98.25/98.47</td>
</tr>
<tr>
<td>HIPAA binary strict</td>
<td>96.50/96.03/96.26</td>
<td>96.67/96.49/96.58</td>
<td>97.73/96.95/97.33</td>
<td>97.28/96.44/96.86</td>
</tr>
</tbody>
</table>

Table 4. Comparison of Bi-LSTM-CRF variants with different neural language models on the CEGS N-GRID 2016 dataset.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Bi-LSTM-CRF (Skip-gram)</th>
<th>Bi-LSTM-CRF (ELMo)</th>
<th>Bi-LSTM-CRF (BERT)</th>
<th>Bi-LSTM-CRF (Skip-gram+LM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P/R/F1(%)</td>
<td>P/R/F1(%)</td>
<td>P/R/F1(%)</td>
<td>P/R/F1(%)</td>
</tr>
<tr>
<td>Token</td>
<td>91.37/91.78/91.57</td>
<td>93.82/91.97/92.89</td>
<td>95.36/92.32/93.81</td>
<td>94.10/91.55/92.81</td>
</tr>
<tr>
<td>Strict</td>
<td>89.58/89.19/89.38</td>
<td>91.37/90.03/90.70</td>
<td>93.39/90.31/91.82</td>
<td>92.19/89.01/90.57</td>
</tr>
<tr>
<td>Relaxed</td>
<td>89.71/89.31/89.51</td>
<td>91.54/90.19/90.86</td>
<td>93.56/90.47/91.99</td>
<td>92.36/89.18/90.74</td>
</tr>
<tr>
<td>Binary token</td>
<td>95.19/95.62/95.4</td>
<td>96.83/94.92/95.87</td>
<td>97.70/94.58/96.11</td>
<td>97.15/94.52/95.81</td>
</tr>
<tr>
<td>Binary strict</td>
<td>92.63/92.22/92.42</td>
<td>93.75/92.37/93.06</td>
<td>95.26/92.11/93.66</td>
<td>94.52/91.26/92.86</td>
</tr>
<tr>
<td>HIPAA token</td>
<td>93.39/93.30/93.34</td>
<td>94.48/93.61/94.04</td>
<td>96.48/93.91/95.18</td>
<td>94.96/93.44/94.20</td>
</tr>
<tr>
<td>HIPAA strict</td>
<td>91.41/91.34/91.37</td>
<td>92.28/91.73/92.00</td>
<td>94.71/92.31/93.50</td>
<td>93.26/91.26/92.25</td>
</tr>
<tr>
<td>HIPAA relaxed</td>
<td>91.58/91.51/91.55</td>
<td>92.49/91.93/92.21</td>
<td>94.92/92.52/93.70</td>
<td>93.44/91.44/92.43</td>
</tr>
<tr>
<td>HIPAA binary token</td>
<td>94.54/94.44/94.49</td>
<td>95.47/94.59/95.03</td>
<td>96.96/94.39/95.66</td>
<td>96.02/94.49/95.25</td>
</tr>
<tr>
<td>HIPAA binary strict</td>
<td>92.35/92.28/92.31</td>
<td>92.95/92.39/92.67</td>
<td>95.13/92.72/93.91</td>
<td>93.97/91.96/92.95</td>
</tr>
</tbody>
</table>

The results of the best Bi-LSTM-CRF variant, Bi-LSTM-CRF(BERT), on the main seven categories are shown in Table 5. On the i2b2 challenge dataset, Bi-LSTM-CRF(BERT) performs best on DATE with a “strict” F1-score of 98.75% and worst on PROFESSION with a “strict” F1-score of 82.78%, while on the CEGS N-GRID 2016 dataset, Bi-LSTM-CRF(BERT) performs best on DATE with a “strict” F1-score of 97.03% and worst on ID with a “strict” F1-score of 59.70%.

Table 5. Detailed results of the best Bi-LSTM-CRF variant on the main seven categories (“strict”).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>i2b2 2014</th>
<th>CEGS N-GRID 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P (%)</td>
<td>R (%)</td>
</tr>
<tr>
<td>NAME</td>
<td>95.26</td>
<td>94.83</td>
</tr>
<tr>
<td>PROFESSION</td>
<td>82.32</td>
<td>83.24</td>
</tr>
<tr>
<td>LOCATION</td>
<td>90.47</td>
<td>86.93</td>
</tr>
<tr>
<td>AGE</td>
<td>97.51</td>
<td>97.38</td>
</tr>
<tr>
<td>DATE</td>
<td>98.97</td>
<td>98.53</td>
</tr>
<tr>
<td>CONTACT</td>
<td>96.36</td>
<td>97.25</td>
</tr>
<tr>
<td>ID</td>
<td>92.97</td>
<td>91.04</td>
</tr>
</tbody>
</table>
We also compare other state-of-the-art methods with Bi-LSTM-CRF(BERT) as shown in Table 6, where only “binary token” P, R and F1 are reported as the P, R and F1 under other criteria are not complete. “Bi-LSTM-CRF(BERT)” shows better performance than the other methods for comparison.

Table 6. Comparison of other state-of-the-art methods with Bi-LSTM-CRF(BERT) (“binary token”).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>12b2 2014</th>
<th>CEGS N-GRID 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>P (%)</td>
<td>R (%)</td>
</tr>
<tr>
<td>Liu et al. [9]</td>
<td>97.94</td>
<td>96.04</td>
</tr>
<tr>
<td>Zhao et al. [14]</td>
<td>98.89</td>
<td>97.23</td>
</tr>
<tr>
<td>Dernoncourt et al. [10]</td>
<td>98.34</td>
<td>98.53</td>
</tr>
<tr>
<td>Kim et al. [15]</td>
<td>99.16</td>
<td>98.06</td>
</tr>
<tr>
<td>Bi-LSTM-CRF(BERT)</td>
<td>99.02</td>
<td>98.38</td>
</tr>
</tbody>
</table>

Discussion

It is easy to understand that Bi-LSTM-CRF(ELMo) and Bi-LSTM-CRF(BERT) perform better than Bi-LSTM-CRF(Skip-gram) since ELMo and BERT can dynamically give more detailed representation of each word according to its context. For example, “Brigham Young” is usually a name, but an organization in “Father is supervisor. College (Brigham Young, community college in TX, UCSD, Baylor), still working on bachelor’s, major sociology.”, which is correctly recognized by Bi-LSTM-CRF(ELMo) and Bi-LSTM-CRF(BERT), but wrongly recognized as a doctor’s name by Bi-LSTM-CRF(Skip-gram). Similar to previous studies in other domains[19], Bi-LSTM-CRF(BERT) shows better performance than Bi-LSTM-CRF(ELMo). As LM has ability to predict sentences regarding a specific task, Bi-LSTM-CRF(Skip-gram+LM) can fit de-identification here.

Although Bi-LSTM-CRF(BERT) shows promising overall performance, there are still some errors. Taking the results on the CEGS N-GRID 2016 dataset for example, the effect of boundary errors or category errors on overall F1-score is around 2.0% (the “strict” F1-score of 91.82% vs the “token” F1-score of 93.81% or the “binary strict” F1-score of 93.66%). The “strict” F1-scores of Bi-LSTM-CRF(BERT) on PROFESSION, LOCATION and ID are much lower than other categories as shown in Table 4. We look into the errors in these three categories and find that: 1) Lots of instances in PROFESSION are not detected as they are not mentioned directly like PHI instances in other categories, for example, “danced” is not a word to denote PROFESSION directly, but we can infer that the patient is a dancer from “She danced with MBT for 10 years;”. 2) Boundary errors widely exist in LOCATION, for example, in “video editing (DiC Entertainment Affiliate) in broadcasting Financial Stress: Yes”, “DiC Entertainment Affiliate” is recognized as an ORGANIZATION instance instead of “DiC Entertainment”. The “binary strict” F1-score on LOCATION is 89.13%, higher than the “strict” F1-score on LOCATION (i.e., 85.33%) by 3.80%. Other than that, there are also more category errors on LOCATION than other categories because of more subcategories in LOCATION. 3) The reason why Bi-LSTM-CRF(BERT) performs bad on ID is that the number of ID instances is too small. From Table 1, we can see there are only 44 ID instances in the training set of the CEGS N-GRID 2016 challenge, in particular, only 2 instances in IDNUM.

For the future work, there may be two directions for further improvement: 1) using BERT pretrained on large-scale clinical data; 2) integrating LM and BERT into Bi-LSTM-CRF at the same time since they work at different layer.

Conclusion

In this study, we introduce Bi-LSTM-CRF with neural language models for de-identification of clinical text. The Bi-LSTM-CRF variants achieve the highest “strict” micro-averaged F1-score of 95.50% on the i2b2 2014 de-identification dataset and 91.82% on the CEGS N-GRID 2016 de-identification dataset, constituting new benchmarks for de-identification.

Acknowledgement

This paper is supported in part by grants: NSFCs (National Natural Science Foundations of China) (U1813215, 61876052 and 61573118), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20160531192358466), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).
References

Achieving Data Liquidity: Lessons Learned from Analysis of 38 Clinical Registries (The Duke-Pew Data Interoperability Project)

Authors: James E. Tcheng, MD\textsuperscript{1}; Joseph P. Drozda Jr., MD\textsuperscript{2}; Davera Gabriel, RN\textsuperscript{1}; Anne Heath\textsuperscript{1}; Rebecca W. Wilgus, RN MSN\textsuperscript{1}; Mary Williams\textsuperscript{1}; Thomas A. Windle\textsuperscript{3}; John R. Windle, MD\textsuperscript{3}.

\textsuperscript{1}Duke Clinical Research Institute, Durham, North Carolina; \textsuperscript{2}Sisters of Mercy Health System, St. Louis Missouri; \textsuperscript{3}University of Nebraska Medical Center, Omaha, Nebraska
Abstract

Background: To assess the current state of clinical data interoperability, we evaluated the use of data standards across 38 large professional society registries.

Methods: The analysis included 4 primary components: 1) environmental scan, 2) abstraction and cross-tabulation of clinical concepts and corresponding data elements from registry case report forms, dictionaries, and / or data models, 3) cross-tabulation of same across national common data models, and 4) specifying data element metadata to achieve native data interoperability.

Results: The registry analysis identified approximately 50 core clinical concepts. None were captured using the same data representation across all registries, and there was little implementation of data standards. To improve technical implementation, we specified 13 key metadata for each concept to be used to achieve data consistency.

Conclusion: The registry community has not benefitted from and does not contribute to interoperability efforts. A common, authoritative process to specify and implement common data elements is greatly needed.
**Introduction and Background:**

Clinical data interoperability, or the capacity to exchange high-quality, clinically relevant information as data from one system to another, has largely failed. Proprietary processes and custom solutions are required at each point of data exchange to extract, transform, and load data from one system to the next. While healthcare ontologies have been developed to support interoperability, these ontologies are positioned as secondary encoding schemas and are not usable natively as clinical vocabularies. While electronic health record (EHR) systems are now ubiquitous, healthcare continues to largely capture clinical information as analog text requiring data abstraction for computational purposes. The transformation and movement of clinical data from one system to another is expensive, labor intensive, and prone to semantic inaccuracy.

The capture of clinical data for submission to registries is illustrative. Clinical registries collect baseline and longitudinal data to evaluate outcomes, assess care performance and processes, and facilitate process improvement in cohorts of patients defined by a disease state, condition or exposure to a medical product. Registries have been positioned as a critical source of real-world evidence to advance clinical and regulatory science. A common source of data for clinical registries is the EHR. However, the submission of data to registries depends largely on the manual abstraction of data from the EHR and manual re-keying of data into registry data collection systems, termed “swivel chair interoperability”.

To assess the current state of interoperability with respect to registries, we evaluated the adoption and use of data standards by 38 national clinical registries in a project termed “Improving Healthcare Data Interoperability” sponsored by the Pew Charitable Trusts. The hypothesis was that data liquidity had not been achieved in the registry domain, and that native data interoperability spanning clinical documentation systems and registry database systems would provide the best pathway to accomplishing data liquidity. Specifically, we evaluated the current status of the adoption of data standards by clinical registries and created technical (database programming) specifications for a set of more than 50 clinical concepts commonly shared across registries. We evaluated these clinical concepts in the context of existing work including the federal Common Clinical Dataset and the draft 2018 US Core Data for Interoperability, HL7 Fast Healthcare Interoperability Resources, the work of standards organizations such as LOINC, and common data models including OMOP, PCORnet, and SENTINEL. The presentation will share the results of this project and identify suggested next steps.

The project had 3 Aims:

1) Evaluate registry case report forms, registry dictionaries, and / or registry data model artifacts from clinical professional society registries representing a minimum of 20 clinical specialties and patient populations, identifying the core (shared) common clinical concepts collected across the majority of registries

2) Compare the data elements used by the registries to capture the core common clinical concepts in the context of the ‘big 5’ healthcare data standards (SNOMED-CT, LOINC, RxNorm, ICD-10, CPT), developing the minimum metadata needed by database developers to natively implement a consistent technical representation of the concepts

3) Develop a roadmap for the transformations needed to achieve native data interoperability of a common clinical dataset across EHR systems, registries, and national data models

**Methods:**

We solicited participation in the project through broadcast communications to registry members of two registry associations (National Quality Registry Network of the PCPI and the Council of Medical Specialty Societies) and personal communications to a small (<10) number of professional societies. Approximately 75 registries were solicited on our behalf by the two associations (we did not manage the mailing lists per se), with our personal appeals overlapping the association-managed solicitations. Participation was voluntary; all who agreed to participate were included in the analysis. All registries were solicited several times. There were no declinations – the only exclusion was for failure to respond to our invitation. Registries agreeing to participate submitted blank registry case report forms (CRFs), data dictionaries, and / or registry data model representations to the Informatics Group of the Duke Clinical Research Institute, Durham, NC. For national data models, artifacts were sourced from publicly
available content in May-June, 2018. All artifacts were anonymized and kept confidential for purposes of analysis and reporting.

The process we followed is illustrated in Figure 1. We first analyzed the registry artifacts. Clinical concepts were tabulated and the corresponding data elements abstracted from the registry artifacts. We pre-selected for concepts that might be expected to be found in multiple registries (e.g., patient demographics, physical exam findings, procedures, medications, major outcomes). We identified those found across multiple registries as the candidate concepts. Data elements unique to a specific disease or procedure were intentionally excluded from the analysis. From this work, we determined that the data elements of interest could be grouped into the following domains: identifiers / demographics, comorbidities, common physical exam findings, procedure information, medications, lab results and patient outcomes. Attributes assigned by each registry to each data element were recorded in a cross-tabulation. Within each domain, the similarities and differences in data elements across registries were evaluated and a ‘match status’ (degree of concordance) was assigned (identical, nearly identical, no match). Identical concordance was defined as an exact match of data element label and permissible values. Nearly identical concordance was defined as representation of the concepts (particularly permissible values) in identical semantics without an exact match of permissible values (e.g., true / false versus yes / no). Similar concordance reflected variability in the representation of the concept and / or permissible values such that complete translation was not possible (e.g., yes / no versus yes / no / not available). No match indicated the inability to completely translate in terms of semantic meaning from one representation to another. Identical and nearly identical were grouped as “concordant” (i.e., semantically interoperable), while any disagreement was classified as “discordant” (i.e., not concordant).

A similar approach was applied to data elements from the federal Common Clinical Dataset along with national data models including the Observational Medical Outcomes Partnership (OMOP), the FDA SENTINEL initiative, and Patient Centered Outcomes Research Network (PCORnet). Again, similarities and differences in data type, semantic meaning, permissible values, etc., were noted and a concordance status was assigned per the definitions above.

With identification of the set of core common clinical concepts, we next developed metadata (e.g., clinical definition, data type, permissible values, coded permissible values, and terminology bindings) for each of the concepts, focusing on only the metadata required to fully qualify each clinical concept for the purposes of building the concepts as interoperable data in databases. To the greatest extent possible, we relied and prioritized published clinical data standards along with the corresponding HL7 FHIR resource for specific metadata content.

We socialized both the Methods and Recommendations via two stakeholder webinars, an in-person meeting held at the Pew Charitable Trusts headquarters in Washington, DC, and an electronic survey (Qualtrics). Feedback from each was incorporated into the final project artifacts. Additional feedback received from direct communications between the project team and stakeholders was also incorporated. Proceedings from the webinars and in-person meeting were analyzed using NVivo (QSR International), summarized, and incorporated where possible. Recordings and transcripts, along with project work products, are available online at: https://dcri.org/registry-data-standards.

**Results:**

Of the registry owners solicited, 38 agreed to participate. There were no declinations (only non-responses). The artifacts (case report forms, data dictionaries, and / or data models) provided by the 38 registry owners were sufficient to perform the analysis across all 38 registries. Approximately 50 data elements were identified as candidates for the comparative evaluation and metadata development. The original intent was to identify clinical concepts common to 50% or more of the participating registries, however, a very limited number of clinical concepts (sex, date of birth, medications, laboratory results) met this criterion. We therefore elected to use a much lower threshold (20%), deriving the following concepts to be explicitly developed: Patient Name, Date of Birth, Sex, Race, Ethnicity, Procedures, Unique Device Identifier, Vital Signs (height, length, weight, blood pressure, pulse), Lab Results (via a separate model), Medications (via a separate model), Care Team (physician only), Smoking Status (via a separate model), Alcohol Use, Substance Use, and Vital Status (Table 1).
Surprisingly few clinical concepts were collected with precisely the same data element representation from any one registry to the next; there were no concepts captured exactly the same way (i.e., concordant) across all 38 registries. Similarly, common data elements in the national common data models were also surprisingly discordant, with the technical (database) representations appearing to be derived from convenience rather than conformance with a particular data standard. An illustrative example is the clinical concept of sex. Of the 21 registries capturing this information, all intended semantically to reflect the biological concept of birth sex (not the social construct of “gender”). Yet one third (n=7) of registries assigned “Gender” or “Patient Gender” as the data element label. Despite a reasonably well-established value set, there were 6 primary variations of the value set across the registries: Male | Female, 3 sub-variations of Male | Female | Unknown, Male | Female | Other | Unknown, Male | Female | Undifferentiated, Male | Female | Unknown/Missing, and unspecified text (Table 2). Another example, tobacco use, is illustrated in Table 3. For this concept, not one of the 11 registries was concordant with another, despite the federal quality measure and resulting value set regarding same, and only 1 of the registries referenced SNOMED.

Most clinical concepts could be captured in a unidimensional manner (i.e., as a label : value pair). For these clinical concepts, 13 key metadata were specified for each data element, with the priority given to metadata with specific relevance to either (or both) clinicians and database developers (Table 4). Metadata describing the context in which data are collected (i.e., past medical history, within specified timeframe of procedure) were not included in this framework as the implications are different for each use case. Priority values for recommended metadata was given to values from predicate work, particularly the ONC USCDI, the NIH National Library of Medicine (NLM) Value Set Authority Center (VSAC), and HL7 FHIR profiles and resources (especially content listed in the FHIR Detailed Description tabs and FHIR Implementation Guides).

Critical discoveries identified during the specification of metadata included that 1) there was no single source or set of sources where all metadata could be identified – developing the metadata for each concept required the use of multiple sources; and 2) often there was not an obvious, single ‘correct’ choice, particularly for reference ontology bindings. Additionally, many common data elements and corresponding allowed values lacked explicit, unambiguous clinical definitions; in fact, the metadata with greatest variability and the least consistency were clinical definitions and definitions of clinical allowed values.

Two clinical concepts required multiple data elements to capture the semantics required by registries: medications and laboratory results. Medication and laboratory data elements were collected by most registries; 78% collected medication data, while 100% captured laboratory results data. Medication data elements from registries were compared to the AHRQ Health Information Technology Standards Panel model and models from HL7, ICD, NCI, LOINC, SNOMED CT, RxNorm (including RxClass), NDF-RT, NDC and UNII. While over 20 classes of medications metadata were identified in the environmental scan, for the representation of medications in registries, only 4 contexts were identified – the need for a complete list of a patient’s medications by name, medications by therapeutic domain, medications by pharmacologic class, and medications administered in the context of a given procedure. Capture of laboratory data also was multidimensional, typically as a series of observations with multiple prompts related to timing and units of measure. Standards contributing to the laboratory result models include LOINC, SNOMED-CT, and UCUM. Proceedings from two federal multi-agency sponsored public workshops were factored into the recommendations of the laboratory results recommendations.

Conclusions:

Collectively, the registry community is not aligned with and does not contribute to interoperability efforts. The inability to exchange data between EHR systems and registries and the inability to share standardized data across registries are barriers to interoperability and achieving the goal of generating real-world evidence. The project further noted that the specification of data elements in federal initiatives such as the USCDI are not oriented toward clinical workflow and utility. A prime example is the representation of tobacco use per SNOMED-CT codes, rather than reflecting how questions about tobacco use are actually prompted in the clinic, used for clinical purposes, or evaluated in the medical literature.

The technical output of the project was a recommended technical implementation of core common clinical data elements for database programmers of electronic health information (including registry) systems. Should all parties
conform to the implementation, data liquidity and native data interoperability for the project concepts will have been achieved. The artifacts of the project are publicly available at www.dcri.org/registry-data-standards. Socialization efforts are underway. Next steps include first the acceptance and implementation by the registry community, and second the execution of projects that demonstrate the interoperable exchange of data between health care providers (via their electronic health information systems) and registry owners.

The work products also include a white paper that lists details of the project, recommendations for federal partners, registry owners, healthcare organizations, and standards development organizations. The environmental scan found that the current landscape lacks a single, authoritative approach to advance interoperability within and across the registry community. Key recommendations include identification of a common authoritative process to identify, define and specify standards for common clinical data elements and an agreed upon process for governance thereof. Also needed is a common data element repository or common clinical data element library to support the technical adoption of standard common data elements. Similarly, common data element, model tooling, and terminology repositories for candidate data elements are needed.

Collectively, we can leverage many opportunities that will elevate our health care ecosystem to one where the capture of clinically relevant data at point of care also serves the needs of care delivery, outcomes evaluation, quality/performance measurement, and medical product evaluation/surveillance. Doing so increases the availability of real-world data to create real-world evidence, contribute to knowledge generation and translate knowledge into practice to improve public health.

References

Acknowledgement

This work was supported by a grant from the Pew Charitable Trusts. The authors wish to acknowledge the many contributions of our Pew collaborators Ben Moscovitch, Josh Rising MD, and Anqi Lu.
Figure 1. Improving Healthcare Data Interoperability Project Process

Table 1. Classes of Common Clinical Concepts Across Registries Modeled

1. Patient name
2. Date of birth
3. Sex
4. Race
5. Ethnicity
6. Procedures
7. Unique Device Identifier (UDI)
8. Vital signs (height, length, weight, blood pressure, pulse)
9. Laboratory results (via a separate model)
10. Medications (via a separate model)
11. Care team: physician
12. Smoking status (via a separate model)
13. Alcohol use
14. Substance use
15. Vital status (death)
Table 2. Example – Concordance Analysis of Clinical Concept: Sex

<table>
<thead>
<tr>
<th>Data Element Name</th>
<th>Allowed (Permissible) Values</th>
<th>Number of Concordant Instances (1=unique)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Undifferentiated</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>1=Female</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2=Male</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3=Unknown</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>Male</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>Sex of patient</td>
<td>[unspecified text]</td>
<td>2</td>
</tr>
<tr>
<td>Patient's Sex at Birth</td>
<td>m= male</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>f= female</td>
<td></td>
</tr>
<tr>
<td></td>
<td>u= unknown</td>
<td></td>
</tr>
<tr>
<td>Sex (at birth)</td>
<td>Male</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>PATIENT SEX</td>
<td>M</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>Patient Gender</td>
<td>Male</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>1=Male</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2=Female</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1=Unknown/missing</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Example – Concordance Analysis of Clinical Concept: Tobacco Use

<table>
<thead>
<tr>
<th>Data Element Name</th>
<th>Allowed (Permissible) Values</th>
<th>Number of Concordant Registries (1=unique)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SMOKING STATUS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tobacco Use</td>
<td>Never smoker</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Current every day smoker</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Current some day smoker</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smoker, current status (frequency) unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Former smoker</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smoking status unknown</td>
<td>1</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td>Never</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Former</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Current - Every Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Current - Some Days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Current - Frequency Unknown</td>
<td>1</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td>Never</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Current</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quit within past 12 months</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quit more than 12 months</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Screening not performed for medical reasons</td>
<td>1</td>
</tr>
<tr>
<td>Smoking status</td>
<td>(i) Current everyday smoker (449868002)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(ii) Current some day smoker (428041000124106)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(iii) Former smoker (8517006)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(iv) Never smoker (266919005)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(v) Smoker, current status unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(vii) Heavy tobacco smoker (428071000124103)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(vi) Unknown if ever smoked (266927001)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(viii) Light tobacco smoker (428061000124105)</td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>0 = Never</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 = Prior</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2 = Current</td>
<td></td>
</tr>
<tr>
<td>Current/Recent smoker (&lt; 1 year)</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>Does the patient currently smoke?</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>If Current or Quit within 12 months, Smoking cessation counseling provided?</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>Tobacco Type</td>
<td>Cigarettes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cigars</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pipe</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smokeless</td>
<td>1</td>
</tr>
<tr>
<td>Types of Nicotine Use</td>
<td>Smoking</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chewing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E-cigarette</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Patch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gum</td>
<td>1</td>
</tr>
<tr>
<td>Tobacco w/in 1 year - Cigarette</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 4. Key Common Clinical Data Element Metadata Needed for Clinical and Database Implementation

1. Clinical concept label (human prompt – for case report form, data entry screen)
2. Clinical definition
3. Clinical allowed values (human prompt – for case report form, data entry screen)
4. Clinical allowed values definitions
5. Database field label
6. Database field data type / format (e.g., char, date, integer, values set)
7. Database field business rules (edit checks, range checks, etc.)
8. Database allowed values (as stored in a database)
9. Object identifier (OID)
10. Reference ontology concept binding
11. Reference ontology allowed values bindings
12. FHIR URL references (profiles, resources)
13. Sources, references, other notes
Predicting Nocturnal Hypoglycemia from Continuous Glucose Monitoring Data with Extended Prediction Horizon

Long Vu, PhD\textsuperscript{1}, Sarah Kefayati, PhD\textsuperscript{2}, Tsuyoshi Idé, PhD\textsuperscript{1}, Venkata Pavuluri, MS\textsuperscript{1}, Gretchen Jackson, MD, PhD\textsuperscript{2}, Lisa Latts, MD, MSPH\textsuperscript{2}, Yuxiang Zhong, PhD\textsuperscript{3}, Pratik Agrawal, MS\textsuperscript{3}, Yuan-chi Chang, PhD\textsuperscript{1}

\textsuperscript{1}IBM Research AI, Yorktown Heights, NY, USA, \textsuperscript{2}IBM Watson Health, Cambridge, MA, USA, \textsuperscript{3}Medtronic, Northridge, CA, USA

Abstract

Nocturnal hypoglycemia is a serious complication of insulin-treated diabetes, which commonly goes undetected. Continuous glucose monitoring (CGM) devices have enabled prediction of impending nocturnal hypoglycemia, however, prior efforts have been limited to a short prediction horizon (~ 30 minutes). To this end, a nocturnal hypoglycemia prediction model with a 6-hour horizon (midnight–6 am) was developed using a random forest machine-learning model based on data from 10,000 users with more than 1 million nights of CGM data. The model demonstrated an overall nighttime hypoglycemia prediction performance of ROC AUC = 0.84, with AUC = 0.90 for early night (midnight–3 am) and AUC = 0.75 for late night (prediction at midnight, looking at 3-6 am window). While instabilities and the absence of late-night blood glucose patterns introduce predictability challenges, this 6-hour horizon model demonstrates good performance in predicting nocturnal hypoglycemia. Additional study and specific patient-specific features will provide refinements that further ensure safe overnight management of glycemia.

Introduction

Hypoglycemia – typically defined as a blood glucose (BG) below <70 mg/dL – is one of the acute complications of diabetes in patients treated with insulin (1). Over half of hypoglycemic episodes occur during nocturnal sleep with a higher prevalence for insulin-dependent diabetes patients that include patients with type 1 diabetes (i.e., insulin dependent) more so than type 2 diabetes (2, 3). Although the definition varies, nocturnal hypoglycemia is usually defined by a predetermined time window for which the international consensus recommendation is from midnight to 6 am, a period that usually includes the duration of nighttime sleep and the longest interprandial interval (1, 4).

Nocturnal hypoglycemia is often asymptomatic due to a sleep-induced effect that shifts the counterregulatory activation to hypoglycemia to a lower threshold(4, 5). Unless severe and followed by a seizure, coma, or other noticeable impairment at the awakening time, it can go unrecognized and be prolonged in patients with type 1 diabetes, and in worst cases, can lead to ‘dead-in-bed’ syndrome (5). Detection of nocturnal hypoglycemia is critically important; if recurrent and undetected, even mild asymptomatic episodes can lead to further impairment of defenses against future episodes due to hypoglycemia-associated autonomic failure and defective counterregulatory responses to subsequent events (4, 6). In addition to the increased risk for future episodes, the daytime effects after an episode of nocturnal hypoglycemia, such as fatigue, impaired mood and higher calorie intake and weight gain, considerably lower quality of life (5).

Several risk factors have been associated with nocturnal hypoglycemia including demographic and non-demographic factors: age, race, BMI, comorbidities, stress, and diabetic nephropathy; physical exercise – with greater risk associated with evening exercise; tight glycemic control; excessive insulin concentration during the day; previous episodes of nocturnal hypoglycemia; and low bedtime glucose levels (2, 7, 8).

With the recent availability of continuous glucose monitoring (CGM) devices, better diabetes management is possible that includes detection of asymptomatic episodes of nocturnal hypoglycemia that go unrecognized with self-monitoring blood glucose (SMBG) devices, which are not commonly used during sleep hours. In fact, the 2017 American Diabetes Association (ADA) consensus statement on measures beyond glycated hemoglobin emphasizes the role of CGM in diabetes management, especially for nighttime and asymptomatic hypoglycemic events, which SMBG devices typically fail to detect, and provides standardized recommendations for analyzing CGM data (1, 9).

The richness of CGM data has enabled extraction of many useful insights, such as retrospective analyses revealing trending patterns, that can increase the daily decision-making power of diabetes patients. Furthermore, CGM data
combined with artificial intelligence (AI) techniques can provide prediction of impending adverse events, such as hyperglycemia or hypoglycemia, within a time horizon leading to the event. Contreras and Vehi provide a review of (10)1849 articles from 2010-2018 applying AI techniques focused on predicting BG levels. The AI techniques were categorized into learning models that included neural networks and decision trees, information discovery models, such as clustering techniques, and models to extract reasoning from information, such as Bayesian methods. The review also reports a smaller group of studies on detection of adverse effects of diabetes, such as hypoglycemia or hyperglycemia. The majority of the reported studies in this category, however, address real-time detection of adverse effects rather than predicting future events within a prediction horizon.

Very few researchers have developed models to predict nocturnal hypoglycemia using CGM data (11-13). One model, for instance, was based on a Kalman filter method with earlier version of the work including voting algorithm approach (11). Another group of researchers used a support vector regression model originally developed for diurnal hypoglycemia detections with added features of physical exercise and sleep time to predict nocturnal hypoglycemic events (12). These studies were limited by small cohorts, dependency on additional sensors (e.g., motion, sleep), and unrealistically optimal patients’ data entry in controlled environments. In addition, a major limitation of these methods is the short prediction horizon, typically chosen at 30 minutes prior to impending hypoglycemic event. With the short prediction horizon, such models are usually used in the context of CGM-enabled insulin pump devices, which have an automatic corrective mechanism to control and suspend the basal insulin delivery. However, the long prediction horizon, such as the work proposed herein, is applicable to broader CGM users including the non-pump users.

Another important factor that has not been examined in the prediction of nocturnal hypoglycemia is the marked distinction between the early (normally defined by midnight-3 am) and late portion of nighttime sleep (3-6 am). The early stage of sleep is considered in the slow wave sleep (SWS) while late sleep is dominated by rapid eye movement (REM) portion of the sleep cycle. These phases of the sleep cycle dictate different glucose needs from the brain, which are markedly reduced in early sleep compared to late night sleep or awake time (14). It is reported that 60% to 70% of nocturnal hypoglycemic episodes occur during late sleep, between 3 am and 7 am (14). Compared to early nighttime hypoglycemia, the counterregulatory hormone response is significantly weaker during late night hypoglycemia (5). Moreover, due to distinct physiological differences, some of the treatment recommendations, such as bedtime snack, were found to be effective for only early nighttime hypoglycemia, but not for late night events (5).

The most ideal approach for minimizing nocturnal hypoglycemia is, therefore, to enable patients to manage their impending hypoglycemia well before going to sleep. This means extending the prediction horizon for the duration of sleep. This manuscript describes an effort to develop a prediction model from CGM data with an extended prediction horizon. The early performance results along with the challenges of early and late-night hypoglycemia prediction are reported.

Methods

Based on the ADA guideline of international consensus on using CGM device (1), nocturnal hypoglycemia was defined by the glucose level falling below 70 mg/dL and lasting at least 15 minutes. An overall nighttime hypoglycemic event was defined by the episode happening from midnight-6 am; early-night hypoglycemia was defined from midnight-3 am, and the late-night events from 3-6 am.

Over 1 million nights/days of CGM data were voluntarily uploaded by 9,800 individuals with type 1 diabetes using the MiniMed™ 530G system (Medtronic MiniMed Inc., Northridge, CA) Data from a 6 pm to 6 am window were analyzed. MiniMed™ 530G system users had the option of turning on the system’s threshold suspend feature and setting it to a value between 60-90 mg/dL in consultation with their physician. Upon suspension of insulin delivery, an alarm alerts the user of the pump suspension and, if not disabled by the user, the suspension can last up to 2 hours before insulin delivery resumes. All of the users in the present study had the threshold suspend feature activated with different predefined low threshold, which for the majority was set to 60-70 mg/dL.

The prediction was based on a random forest machine learning model composed of 100 trees and built using python scikit-learn library. Each decision tree spits out a prediction score and the final class is defined by the class with the highest ensemble-average probability score. The random forest classifier was selected for its robustness, accuracy, scalability, and ease of maintenance. For the testing purposes, prior to model finalization, a smaller dataset comprised of approximately 1000 users each with an average of 300 nights of uploaded data (total of 300K nights) was also considered. The positive label for the nights in the dataset was assigned if at least one nocturnal hypoglycemic event was detected per night. The total night samples were first sorted from the furthest night to the most recent night in increasing order by the timestamp of the users streaming the data. The first 80% of the dataset was then considered
for training (10-fold cross validation) and the latest 20% of the data was reserved for the holdout test. The model performance was assessed based on the night data in the test set using the area under the curve (AUC) of the receiver operating characteristics (ROC) curve.

The feature vector was composed of three categories of features: SG features, count-based features, and temporal location features. SG features were extracted as statistical properties of the previous sensor glucose time series. Count-based features were simply a count of the number of hypoglycemic events in the past windows. Temporal location features captured the time of day and day of week, which are important because hypoglycemic events have a tendency to recur on a daily or weekly basis. For the same model, three types of performance for nocturnal hypoglycemia prediction were assessed with similar feature vectors and prediction timepoint at midnight, which included 1) overall nighttime with 6-hour prediction horizon (midnight-6 am); 2) early-night with 3-hour prediction horizon (3-6 am); and 3) late-night with 6-hour prediction horizon (midnight-6 am) only looking to predict if the hypoglycemic event happened 3-6 am.

Results

Study Population

The 9,800 users had a mean±SD age of 45.34±16.38 years (51% female and 49% male) and had been using the MiniMed™ 530G device on average for about 100 days and had been on insulin at least 1 year and maximum of 71 years. The CGM device enabled users to log their carbohydrate consumption as well as administer insulin boluses, although these auxiliary inputs were at the users’ discretion and not consistent throughout their uploads. The prevalence of nocturnal hypoglycemia among the users was 18.7% for the overall nighttime period (12-6 am), where 12.7% of the users exhibited early-night and 11.8% exhibited late-night hypoglycemia.

Overall Nighttime Prediction

The overall nighttime prediction covered the prediction horizon of 6 hours, from the midnight to 6 am, for each user. This prediction duration was for any hypoglycemic event that occurred regardless of its occurrence during early or late night. The overall performance of the model based on the total number of nights (collected from different users) in the test dataset was 0.84 as depicted in Figure 1 by the full AUC metric. The scatter plots show the AUC-based model performance for each user in the test dataset as a function of hypoglycemia prevalence in the totaled nights (about 300 nights) for each user. When the number of users was increased 10 fold, the performance of the model remained unchanged (Figure 1b). The performance of the model deviated less from the mean as the prevalence of hypoglycemia increased indicating more stable hypoglycemia detection for users with higher recurrence.

![Figure 1: Model performance for the overall nighttime hypoglycemia prediction with 6-hour prediction horizon for approximately 1000 users (a) and 9800 users (b). Scatter points represent AUC values for each user in the test dataset. Full AUC represents the value calculated over all the data points included in the test set.](image)

Early versus Late-Night Prediction

The counterregulatory hormone response to low BG during the early and late portion of the night makes differs for these two stages of sleep with respect to predictability of hypoglycemic occurrence. For the early-night hypoglycemia
prediction with prediction horizon of 3 hours from midnight, the model yields 0.90 AUC-based performance as calculated by the full AUC metric (Figure 2a). A separate prediction for the late night for the same user population, however, reveals a lower performance of 0.75 by the model (Figure 2b). For both predictions and for the users with greater hypoglycemia prevalence (referred to as skew in the plots), the user-based performance of the model approached the mean value.

![Figure 2: Model performance for early-night model with 3-hour prediction horizon (a) and late-night model with 6-hour prediction horizon (b) for 9800 users. Scatter points represent AUC values for each user in the test dataset. The full AUC represents the value calculated over all the data points included in the test-set.](image)

**Late-Night Patterns**

To understand the more specific patterns associated with the late-night hypoglycemia, the smaller dataset of about 1000 users (approximately 300 thousand nights of data) was assessed more closely. The variable-wise probability density functions of two samples were compared for the same components of the feature vector as described in the Methods section: the sample size of 11% (of all the nights) with history of late-night hypoglycemia (positive labels) and the sample size of 77% with no late-night history of hypoglycemia (negative labels).

The comparison revealed some unique characteristics for user with late-night hypoglycemia which include: lower SG values at midnight, average downward SG trend before midnight, higher recent hypoglycemic event count, and, in general, a non-zero count of late-night hypoglycemia over the past couple of months.

**Trends in False Negatives Predictions of Late-night Hypoglycemia**

Comparing the early-night and late-night predictions, the rate of false positives was similar, but the false negatives were significantly higher in the late-night predictions. For the late-night false negative predictions, the users could be categorized into four distinct classes as described in Table 1. The illustration of an example for each class is shown in Figure 3 for class 1 (Figure 3a), class 2 (Figure 3b), class 3 (Figure 3c), and class 4 (Figure 3d).

<table>
<thead>
<tr>
<th>False-Negative Classes</th>
<th>Description</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class #1</td>
<td>SG measurement above 180 mg/dL at midnight, sliding below 70 mg/dL between 3 am and 6 am</td>
<td>45%</td>
</tr>
<tr>
<td>Class #2</td>
<td>SG trajectory largely outside of 1-σ band from recent nights</td>
<td>56%</td>
</tr>
<tr>
<td>Class #3</td>
<td>SG trajectory exhibited U or W shaped bounce in the middle of night</td>
<td>33%</td>
</tr>
<tr>
<td>Class #4</td>
<td>SG trajectory not fitting above categories</td>
<td>15%</td>
</tr>
</tbody>
</table>
Discussion

This manuscript describes one of the first attempts to use CGM data to predict nocturnal hypoglycemia with an extended time horizon on a large cohort (~10,000 individuals) with more than 1 million nights of sample data from a wide demographic population (Insert some of the demographic attributes). The main goal of this study was to assess the feasibility of predicting nocturnal hypoglycemia solely based on CGM sensor glucose readings. The presented model can predict hypoglycemia events throughout the nocturnal sleep interval with the prediction horizon of 6 hours at the midnight timepoint with an AUC of 0.84. One previously reported predictive model using a prediction horizon greater than one hour showed that performance decreased with horizon extension, but was optimized when set on 30 minutes (15). To our knowledge, this is the first study using CGM data with a 6-hour extended prediction horizon and high performance, which might enable patients to better address potential hypoglycemia prior to going to sleep.

The main limitation of models with short prediction horizons that are typically around 30 minutes is that they are normally coupled with an automatic insulin suspension (shut-off). This automatic control was designed to reduce hypoglycemic exposure and to help individuals prone to hypoglycemia and/or with impaired hypoglycemia awareness (14). Although, overall, predictive low glucose suspension has been shown to effectively reduce exposure and frequency to nocturnal hypoglycemia (16-18), it can result in some elevation of glucose levels during the night and in the morning, or increase the time spent in moderate hyperglycemia (17-19).

It is also important to note that the use of threshold suspend is not particularly designed to eliminate the hypoglycemic events since it normally becomes active at the start point of the episode; however, depending on the threshold setting, this feature can be helpful in shortening a hypoglycemic event. As a result, the use of threshold suspend in our subjects did not eliminate the nocturnal hypoglycemic episodes. Similar observation has been reported by the Juvenile Diabetes...
Research Foundation study (6) where frequent and prolonged hypoglycemic episode were noticed despite the activated nighttime CGM profile to adjust overnight basal delivery rate. Thus, the ultimate objective of this research is to benefit large spectrum of CGM users by leveraging the large data set collected from insulin pump users to evaluate the potential of nocturnal hypoglycemia prediction.

**Challenges of Late-Night Hypoglycemia Prediction**

Due to differences in counterregulatory responses to low-glucose levels during the early portion of the sleep cycle versus the late portion (dominated by REM sleep), as well as influences from circadian factors, it is expected that hypoglycemia prediction follows a different pattern in each portion of sleep. In fact, this study showed that, when using the same feature parameters, prediction of early-night hypoglycemia results in significantly a higher performance (AUC = 0.90) compared to late-night prediction (AUC = 0.75). While the false-positive rate was similar for both predictions, late-night prediction suffered from significantly higher false negatives and thus, a greater chance of missing an impending late-night hypoglycemic episode.

This observation indicates that a different or additional set of features that can more specifically relate to late-night sleep characteristics may be required for improved prediction. More specific features, in turn, implies having thorough understanding of the relationship between the late-night portion of sleep and hormonal response to BG regulations. Unfortunately, for insulin-dependent patients, this relationship is not fully understood as the bolus insulin and body’s natural insulin secretion do not follow similar glucose regulation (2). In fact, when it comes to insulin and body counterregulatory responses, contradictory findings have been reported. For example, the Somogyi effect or rebound hyperglycemia, is the phenomenon that causes the BG to rise to the hyperglycemic level in early morning in counterreaction response to a severe drop in late-night BG to hypoglycemic levels. This Somogyi effect could perhaps be recognized in some of subjects in this study classified as #1 and # 3 in Table 1. However, some other researchers have challenged the Somogyi effect and in fact, reasoned a lower rather than higher BG level in the morning due to sleep-induced impaired counterregulatory responses (5). Adding to the complexity of the matter is the fear of encountering a hyperglycemia episode in patients that face an elevated SG reading at bedtime; this result could motivate them to administer bolus insulin or skip the valuable bed time snack. For instance, in our cohort, 30% of the nights SG readings started above 180 at midnight; subsequently 6% of those nights had late-night hypoglycemia. This adds to the challenges of predicting impending hypoglycemia solely based on SG readings if at the prediction timepoint an SG reading is at a hyperglycemic range.

Although several risk factors have been associated with nocturnal hypoglycemia, the predictive power of those risk factors still remains to be fully investigated and trusted (5). Some researchers have found weak link between patients demographics, such as age, sex, years of diabetes, and hypoglycemia (20). While the Somogyi effect seems to be correlated with a higher sum of bolus insulin, Woodward et al. (21) found neither an association of total insulin bolus to nocturnal hypoglycemia, nor any link between type of insulin, and frequency of insulin injection. The only predictive factor reported was bedtime SG level.

Besides patients’ demographic and CGM-based SG trajectory measurements, other studies on nocturnal hypoglycemia have looked to include more auxiliary input data provided by the patients to strengthen the prediction power. These additional data inputs include nutritional intake logs and insulin regimen (type, dose, time). Although valuable inputs, and possible to collect under controlled study conditions, the lack of consistency of log entries introduces a challenge for building a predictive model in free-living situations.

In addition to nutritional intake and insulin regimen, physical exercise and emotional state (such as stress and anxiety) have shown to affect glucose hemostasis. For instance, it has been reported that daily stress can increase the instability of glucose levels and risk of hypoglycemia (22). Physical exercise for insulin-dependent diabetes patients has been among the factors most strongly associated with severe hypoglycemia. Exercise enhances the insulin sensitivity in the muscles which can result in lowering of BG levels up to 8-10 hours after exercise. Thus, if performed during the evening, strenuous activity can pose a great risk for late-night hypoglycemia (23). In fact, Campbell et al. (7) reported that the risk of hypoglycemia following an evening exercise can be alleviated by consuming the snack only for the early stage of sleep but not in the late portion of the night. Due to the strong association of exercise and emotional state of the patients with insulin and glucose metabolism, some studies have used additional sensors to track these stimuli and include them in the feature space of the predictive models (12, 13). Eren-Oruklu et al. (13) showed improved performance of their hypoglycemia predictive model with inclusion of emotion and physical activity compared to their previously developed model without these factors. Although important and beneficial, the inclusion
of these factors and the use of additional devices besides CGM can also introduce a challenge in predicting late-night hypoglycemia in free-living conditions.

Besides exercise and stress, it is important to recognize the impact of the sleep pattern on the body’s physiological hormonal response to BG regulation. According to the report by the Centers for Disease Control and Prevention, about 1 in 3 adults do not get enough sleep and 15% of Americans work the night shift and many more work at night in addition to their day work thus, having irregular circadian rhythm (24). Irregular sleep pattern, short duration of sleep, and poor-quality sleep have been associated with prevalence of diabetes and impaired insulin response, and directly impact the nocturnal counterregulatory response to hypoglycemia (25). Although a predefined window of sleep from midnight to 6 am was considered in this study according to the recommendation by International Consensus on Use of CGM (1), it is important to recognize that not all the subjects sleep during this window and for the full duration.

Future work will focus mostly on the nocturnal hypoglycemia and on developing methods to include more patient input, most importantly emotion, exercise, and sleep pattern. This will also include methods to detect meals consistently since food logs could be inconsistent. Moreover, since some of the false-negative cases revealed some of the late-night glucose profile characteristics, such as Somogyi effect, we anticipate prior detection of these events could also help with the prediction modeling. Some of the long-term effort will also focus on developing personalized models to help patients better manage their diabetes according to their living conditions, such as their unique sleep pattern and BG profile.

**Conclusion**

This manuscript describes a model using CGM data to predict nocturnal hypoglycemia with acceptable performance with an extended prediction horizon. Early night prediction was more accurate than late night prediction. We attribute challenges in late-night hypoglycemia prediction mainly to lack of clinical understanding of some of the residual effects during the day manifesting at this portion of the sleep. We anticipate inclusion of additional BG-altering stimuli, such as emotion, exercise, and sleep patterns, that may improve late-night hypoglycemia predictability. However, inclusion will require additional devices that may not be convenient for patients in free-living conditions. Future work will focus on more data-driven methods to predict patient-specific inputs in order to enrich the model feature vector and improve the predictability.

**References**


Using Electronic Health Record Activity to Represent Interdisciplinary Care Teams and Examining their Contribution to Hospital Length of Stay

Dammika L. Walpitage, Ph. D1, Amy Garcia, DNP, RN, CENP2, Ellen Harper DNP, RN-BC, MBA, FAAN2, Neena K. Sharma, PT, PhD3, Lemuel R. Waitman, Ph. D1,
1Division of Medical Informatics, Department of Internal Medicine, 2School of Nursing, 3Department of Physical Therapy and Rehabilitation Science, University of Kansas Medical Center

Abstract
Modeling variance in patient outcomes using medical claims and other forms of aggregated administrative data may ignore significant contributions associated with providers who are not recorded in billing transactions. We examined the association between interdisciplinary provider factors and length of stay (LOS) for 1,099 lumbar spine surgery patients. Interdisciplinary provider “dose” (number of providers/case), “workload” (care of other patients), and “activity” factors were defined and generated. Hierarchical Regression models were used to test the impact of these provider factors controlling for the effect of socio-demographic and clinical factors. Interdisciplinary provider factors explained 12% of additional variance in LOS. EHR-based interdisciplinary care team representations hold promise in contributing to our understanding of health care delivery and quality.

Keywords: interdisciplinary care, nursing documentation, workload, length of stay, electronic health records (EHR)

Introduction
While traditional health care delivery and quality analysis incorporates clinical factors1-4 such as disease type and severity, there is considerable variation in outcome that is not explained leading to studies that incorporate additional social determinants of health, environmental, and genetic attributes. Health care delivery systems and provider characteristics may also be considered as a source of such additional variance4-6. Provider characteristics such as licensure, clinical experience, professional education and training, and knowledge of practice guidelines have led to variations in treatment, outcomes, as well as the cost of care1,7,8. Past research that modeled the shared variance between provider characteristics and patient outcomes has been based on billing data, unit level aggregated data, or survey data9-13. However, billing and aggregated data does not represent many of the providers associated with treating a patient, such as nurses, medical residents and other non-physician professional providers. Thus, such models could either mask or combine the contribution of important members of the care team. Furthermore, assessing the contribution of the individual provider, especially nurses, to clinical outcomes is considered to be challenging due to the lack of valid measures14,15. Compared to billing and other forms of aggregated data, electronic health records (EHRs) provide comprehensive clinical and care process data, time-stamped with contribution of individual providers (non-billing)16-18. However, only limited research has utilized EHR data to measure the contribution of individual providers and investigate the impact of provider characteristics on patient outcomes19.

The length of stay (LOS) is a frequently used patient outcome and a proxy measure of the cost that depends on many clinical and non-clinical factors. A comprehensive systematic review of 107 articles on the use of EHR to develop risk prediction models from 2009-2014 only reported three studies with LOS predictions20. However, none of these models include provider attributes, especially related to nursing or other non-physician professionals. There are many studies that
utilized the aggregated and other forms of provider (mainly nursing) data which showed positive effects of provider characteristics on reducing LOS\textsuperscript{21,22}. Results of research reviewing 17 articles examining the relationships between nurse staffing and LOS showed evidence for reductions in LOS associated with higher ratios of nursing personnel in hospital settings\textsuperscript{22}. Furthermore, the recent studies by Cho et.al. (2017) concludes that the nurse staffing (fewer patients per nurse) was significantly associated with the reducing LOS of surgical patients\textsuperscript{23}. Also, a meta-analysis on nursing workforce characteristics by Kane et.al. (2007) found strong evidence on direct care RN hours leading to shorter LOS\textsuperscript{24}. However, measures used in these studies could not distinguish between providers in direct patient care or those involved in administrative or outpatient activities due to the aggregate nature of the data. Also, the cross-sectional datasets used in these studies do not allow the opportunity to study trends or estimate lagged-timing effects. Understanding these trends or lagged effects could better explain the relationship between nursing and other professional care flow and resulting LOS, that could ultimately lead to optimizing the available provider resources for better patient care\textsuperscript{24}.

Factors effecting the LOS among clinical populations managed by interdisciplinary providers is of a greater interest. Populations undergone spinal procedures are generally associated with intense pain in the postoperative period, especially for the initial few days\textsuperscript{25,26}, and are typically managed by interdisciplinary provider teams. Adequate pain management in this period has been seen to correlate well with improved functional outcome and LOS\textsuperscript{25,27,28}. Lumbar Spine surgery (LSS) is a common spinal procedure done in hospitals, and usually considered a treatment of certain spinal disorders when conservative management of pain and other symptoms fail. The rate of LSS is highest in the United States compared to other countries, and associated with high surgery cost and related post-surgical care\textsuperscript{28,29}. Following surgical procedures like LSS, providers from many disciplines work together to reduce the pain and increase mobility for timely discharge. Nurses contribute with ongoing pain assessment, developing the pain management plan, implementing non-pharmacologic interventions (cold therapy, music therapy), and evaluating responses to such interventions for adjusting the pain management plan\textsuperscript{30,31}. Physicians contribute with ordering analgesics (particularly opioids) and ordering physical therapy interventions depending on the patient recovery. The physical therapist (PT) contributes to patient care by assessments and conducting mobility trainings\textsuperscript{28,29}. Respiratory therapists (RT) may also be involved in the care of patient recovering after LSS\textsuperscript{32}.

The purpose of this preliminary cohort study was to examine the effect of interdisciplinary provider factors on the LOS after LSS. We hypothesized that the number of providers (provider dose) performing patient care activities during the first two days following LSS is associated with LOS. Also, we hypothesized that the provider workload across multiple patients under their care influences LOS. Thus, the specific objectives of this study were to: (i) Examine the variance in LOS explained by the provider workload (as defined by their documented activities for other patients during the first two days) above and beyond the variance explained by the socio-demographic and clinical factors for different provider types, (ii) Examine the impact of provider dose (e.g. the number of distinct providers during the first two days) on LOS controlling for the effect of providers’ level of activity within the encounter and their workload on other patients, (iii) Examine the variance in LOS explained by the interdisciplinary provider factors: workload, level of activity, and dose (during the first two days) above and beyond the variance explained by socio-demographic and clinical factors.
Methods

Data Sources and Study Cohort
Data for this retrospective study were extracted from the University of Kansas Medical Center’s Healthcare Enterprise Repository for Ontological Narration (HERON), an i2b2-based clinical integrated data repository\textsuperscript{33,34} that includes EHR, billing records, and clinical registries. We identified the cohort of interest from the HERON system using the i2b2 query and analysis tool with specific current procedural terminology codes to select patients who had posterior LSS consisting of laminotomy, laminectomy, or fusion (arthrodesis) during the period January 2008 to December 2018. Patients 45 years or older with LOS of three days or more (as defined by billing records reported to the Vizient Clinical Database\textsuperscript{35}) and who have inpatient PT assessment data were included. Medical records were excluded for patients with a history of neoplasm, intraspinal abscess, spinal deformity (i.e., scoliosis and kyphoscoliosis), spine fractures, vertebroplasty, osteomyelitis, and cauda equina syndrome. Based on these criteria, a cohort of 1,099 patients was identified.

Selection of Measures and Covariates
All the variables of interest for the study were selected from the HERON system. Clinical and socio-demographics were selected based on the most commonly used variables to predict LOS flowing inpatient stay in general, but are acknowledged to be limited in scope as the focus of this study was to develop and evaluate novel representations of the interdisciplinary care teams\textsuperscript{36,37,38,39}. Interdisciplinary provider factors were generated from the discrete care activities performed by different provider types: medication administrations (by nurses), flowsheet documentation (by nurses), medication orders (by physicians), procedure orders (by physicians), flowsheet documentation (by physical therapists), and flowsheet documentation (by respiratory therapists).

Provider factors described as Dose, Workload, and Activity were defined as follows:

- **Dose**: the number of distinct providers involved with patient care during the first two days. (e.g. 3 distinct nurses administered medications during the first two days). In the data set we have 6 variables, each representing one provider type related activity.
- **Workload**: represents the activities for and severity of the other patients a patient’s providers cared for during the first two days. Workload factors were created using following three sub-components.
  - **Other patients**: number of other patients cared for by the providers (during the two days these 3 nurses did medication administrations for 7 other patients).
  - **Other patients’ severity**: average severity of other patients (average value of All-Patient Refined Diagnosis Related Group (APR-DRG) severity of 7 other patients is 2, which indicates on average they cared moderately severe patients).
  - **Other patients’ activity**: average number of activity (EHR facts available) for other patients (on average these 3 nurses did 15 medications administrations per other patients they cared during the first two days).

- **Activity**: Number of activities (EHR facts available for medication administration by nurse, flowsheet documentation by physical therapists, etc.) for the given patient/encounter (during the first two days these 3 nurses performed twelve medication administration on the given patient).

Abbreviated but representative Clinical and Socio-demographic factors used to establish a baseline for variance analysis included Medical Severity Diagnosed Related Group (MSDRG), Charlson comorbidity index, Intensive Care Units (ICU) days, Age at Admission, Gender, Race.
Data Management and Statistical Analysis

Prior to conducting inferential statistical analysis, univariate normality of continuous variables was tested, and transformations were performed whenever needed (natural log transformation for LOS). Sixteen outliers were identified using criteria of standardized values falling outside the range of -3 to +3 interval\cite{40,41} and removed prior to the analysis. To decompose the variance attributed to various factors, we fitted a hierarchical regression model with natural log transformed LOS as the dependent variable. Variables representing socio-demographic factors and clinical factor were entered at the first step. We updated the baseline model (at the first step) for each provider type specific activity (e.g. medication administration by a nurse, medication orders by a physician, etc.). At the second step we entered the workload factor (to address objective i), and at the third step we entered the provider activity factor (control variable). At the fourth and final steps we entered the provider dose factor (to address objective ii). The variation attributed with each step of the hierarchical model (all provider types combined) was graphically analyzed using a Venn diagram representation (to address objective iii).

Results

The study sample of 1,099 patients had total of 1,146 encounters following LSS with 583 encounters (50.87%) for men and 563 (49.13%) for women. Majority of the encounters 967 (84.38%) were for white patients, while 103 (8.99%) encounters were for African Americans. The average age of these patients at the time of the encounter is 64.94 (SD=12.76). Almost equal proportions of patients in the sample were categorized with moderate severity (43.80%) or minor severity (43.45%), 10.38% of patients in the sample were categorized as having major severity, and remaining (approximately 2%) as extreme severity. Average LOS following LSS for this sample was 5.38 (SD=4.90). 150 encounters (13.09%) had ICU days with mean ICU days being 2.85 (SD=4.79). The average value of Charlson comorbidity index for LSS patients in this sample was 1.98 (SD=1.48).

Providers Associated with Patient Care

During the first two days following LSS, out of the total 223 physicians involved, 135 physicians ordered medications, while 211 physicians ordered medical procedures. Out of the total 1,387 nurses involved, 1,137 nurses administered medication, and 1,058 nurses have documented in flowsheets. Based on the flowsheet documentation there were 76 physical therapists and 183 respiratory therapists treating LSS patients during the first two days following LSS. Summary statistics in Table 1 shows, during the first two days, on average 5 nurses were involved with LSS patients care, while 1 to 2 physicians were involved with ordering medication or procedures. Also, most of the encounters had two physical therapists and two repository therapists involved. The average number of medication orders by physicians (32.66) slightly exceeded the average number of medication administrations (30.99) by nurses; perhaps scheduled orders for administration after the first two days or at discharge. Nursing flowsheet documentation activity was five to six times greater than physical therapists and with higher variance (coefficient of variance; 0.84) between encounters compared to physical therapy documentation activity (coefficient of variance; 0.61).

LOS Distribution for MSDRGs

Patients in the sample were assigned into 34 different MSDRG categories with the majority (40.83%) assigned into MSDRG 460 (Spinal fusion except for cervical without MCC), and 7.15% without any particular DRG code assigned. We combined the other DRGs into meaningful groups by keeping adjacent MSDRGs together. To visualize we recoded the provider dose (number of providers associated with care) for each encounter as “Low” if the provider dose for that particular
encounter was less than the median for all the encounters, and otherwise as “High”. As shown in Figure 1, the encounters assigned with MSDRG:460 having either low or high provider dosage has approximately equal median LOS, with low dose having relatively more outlying cases for all six provider related care activities. For most of the other DRGs, procedure orders by more physicians were associated with longer LOS. Furthermore, for most DRGs flowsheet documentation activity by less nurses, physical therapists, and respiratory therapists was associated with longer LOS.

Table 1: Mean (standard deviation) of interdisciplinary provider factors derived from two days EHR activity

<table>
<thead>
<tr>
<th></th>
<th>Workload (for other patients)</th>
<th>Activity (on patient by all providers)</th>
<th>Dose (number of providers)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Patients</td>
<td>Patient Severity</td>
<td>Activity (per provider)</td>
</tr>
<tr>
<td>Nurse</td>
<td>16.90 (7.07)</td>
<td>2.45 (0.42)</td>
<td>10.70 (3.79)</td>
</tr>
<tr>
<td>Medication Administrations</td>
<td>18.11 (13.06)</td>
<td>2.56 (0.48)</td>
<td>97.03 (78.83)</td>
</tr>
<tr>
<td>Flowsheets Documentations</td>
<td>6.30 (5.03)</td>
<td>2.23 (0.61)</td>
<td>4.93 (3.78)</td>
</tr>
<tr>
<td>Physician</td>
<td>19.49 (15.23)</td>
<td>2.24 (0.52)</td>
<td>14.92 (12.94)</td>
</tr>
<tr>
<td>Medication Orders</td>
<td>76.74 (32.70)</td>
<td>2.82 (0.21)</td>
<td>16.07 (10.23)</td>
</tr>
<tr>
<td>Procedure Orders</td>
<td>48.34 (20.55)</td>
<td>2.78 (0.25)</td>
<td>14.73 (9.76)</td>
</tr>
<tr>
<td>Physical Therapist</td>
<td>48.34 (20.55)</td>
<td>2.78 (0.25)</td>
<td>14.73 (9.76)</td>
</tr>
<tr>
<td>Respiratory Therapist</td>
<td>48.34 (20.55)</td>
<td>2.78 (0.25)</td>
<td>14.73 (9.76)</td>
</tr>
</tbody>
</table>

Note: Severity measured as 1-Minor, 2-Moderate, 3-Major, 4-Extreme.

Figure 1: Relationship between provider dose and LOS (for MSDRGS)

A-Medication administrations by nurses, B-Flowsheet documentations by nurses, C-Medication orders by physicians, D-Procedure orders by physician, E-Flowsheet documentations by physical therapists, F-Flowsheet documentations by respiratory therapists
Decomposing Variance: Baseline Model

Baseline regression model explained 20.39% of variance in LOS with (R=.45, Adj. R² =19.39, F(14,1113)=20.36, p<.001). Race, Age at visit, ICU days, MSDRG and Charlson Comorbidity were significant predictors of LOS (p<.01, p<.05, p<.001, p<.05 respectively). Compared to White patients African American patients (β = .0098 and p<.01) had higher LOS. Compared to encounters with unassigned MSDG, encounters having other DRGs, groups;453-455 (β = 0.163); 456-459 (β = 0.176); 471-473 (β = 0.180) had higher LOS. Age at visit showed a quadratic relationship (β = .000048) with LOS. As ICU days increased by one unit and Charlson comorbidity increased by one unit LOS (at log scale) was incised by (β = .141), and (β = .018) units respectively.

Decomposing Variance: By Different Provider Type and Related Care Activity

The additional variance in LOS that is attributed to each step (for provider factors, beyond the first step) of the hierarchical regression and unstandardized β value for the provider dose effect is shown in Table 2. Variance portioning at the second step showed significant impact of workload factor on explaining variance in LOS among all three non-physician provider related activities, significant impact for medication orders by physicians (p<.05), while non-significant impact for procedure orders by physicians. With respect to nursing, workload factor with flowsheet documentation had slightly higher impact on LOS than medication administration (2.71%, p<.001 compared to 1.98%, p<.01). Variance portioning at the third step showed significant impact of provider activity factor on explaining the variance in LOS only for physicians and physical therapist. Additional variance in LOS explained by the provider dose factor at the fourth step was significant only for physicians and respiratory therapists. Thus, above and beyond the variance explained by all other factors, physician dose and respiratory therapists dose has an impact on explaining additional variance in LOS. Among the set of final four step models, the highest additional variance in LOS was observed for the model with flowsheet documentations by physical therapists. Positive unstandardized β values for provider dose indicates that, controlling other factors, increasing the provider dose had led to increase the LOS in all provider type and related activities. However, this effect is significant only with both types of orders by physicians and respiratory therapy flowsheet documentation activities.

Table 2 - Variance explained by provider factors and provider dose effects

<table>
<thead>
<tr>
<th></th>
<th>Step 2 (Workload)</th>
<th>Step 3 (Activity)</th>
<th>Step 4 (Dose)</th>
<th>Final Model β (for Dose)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medication Administration</td>
<td>1.98***</td>
<td>0.00*</td>
<td>0.01</td>
<td>1.99*</td>
</tr>
<tr>
<td>Flowsheets Documentation</td>
<td>2.71***</td>
<td>0.16</td>
<td>0.15</td>
<td>3.02***</td>
</tr>
<tr>
<td>Physician</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medication Orders</td>
<td>0.7*</td>
<td>1.12***</td>
<td>0.53***</td>
<td>2.35***</td>
</tr>
<tr>
<td>Procedure Orders</td>
<td>0.11</td>
<td>0.57*</td>
<td>3.18***</td>
<td>3.86***</td>
</tr>
<tr>
<td>Physical Therapist</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flowsheets Documentation</td>
<td>2.54***</td>
<td>1.46***</td>
<td>0.00**</td>
<td>4.00***</td>
</tr>
<tr>
<td>Respiratory Therapist</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flowsheets Documentation</td>
<td>2.04***</td>
<td>0.08</td>
<td>0.46*</td>
<td>2.59***</td>
</tr>
</tbody>
</table>

Note: *p<.05, **p<.01, ***p<.001, Low sum of square difference: '00015958, '000076896
**Decomposing Variance: For All Provider Types and Related Care Activities**

Following similar hierarchical approach but using the full set of variables (all 6 workload variables, all 6 activity variables, and all 6 dose variables) for all provider type and related activities, we investigated the overall variability in LOS explained by the provider factors above and beyond socio-demographic and clinical factors. As illustrated in Figure 2, we started with the baseline model which explained 20.39% variance in LOS, and addition of workload factors (now workload for medication administration by nurses, workload for flowsheets documentation by nurses, workload for medication order by physician, etc.) explained 6.40% of additional variance (p<.001) in LOS. Adding provider activity factors explained additional 2.27% (p<.001) of variance in LOS. Adding provider dose at the final step explained additional 3.27% (p<.001) of variance in LOS. Final model with socio-demographic and clinical factors, provider workload, provider activity, and provider dose accounted for 32.33% variance in LOS.

![Figure 2: Variance explained by socio-demographic, clinical, and provider factors](image)

**Discussion**

Our study sought to explore new representations of interdisciplinary care team activity and staffing, and then examined the variance in LOS after Lumbar Spine Surgery attributed to these provider factors. One of the important features of this study was that provider factors used in the analysis have rarely been considered in most of the studies predicting LOS. Further, we have considered modeling the effects of interdisciplinary providers, especially including non-physician providers, as they are extensively involved in patient care. Using comprehensive, time-stamped EHR data we created variables to represent providers’ workload on other patients they cared for during the same day, their activity with patients, and the exact number of providers associated with the care for a given time frame (provider dose). Compared to aggregate data, such variables (factors) provide more accurate and direct measures to investigate the effects of provider attributes on LOS. With newly constructed factors and complied EHR data, we were able to quantify interdisciplinary provider activity following LSS at a more detailed level. Post LSS patient care for the study cohort was performed by interdisciplinary care team consisting of but not limited to, nurses, physicians, physical therapists and respiratory therapists. As a major finding, our study confirmed the significant contribution of interdisciplinary provider factors on explaining the variation in LOS.
While a comprehensive clinical analysis of LOS was not the focus of this study, consistent with previous LOS studies, we found significant contributions of age, race, and ICU days, comorbidities, and medical severity to the variance in LOS for patient stays following LSS. Our results showed that workload on other patients was a more significant factor explaining the variance in LOS than individual patient care activity and staffing. The impact of other patients’ workload on LOS was relatively strong for all three non-physician providers compared to physicians, while nurses’ flowsheet documentation workload showing the most significant contribution on explaining the variance in LOS. Our models identified the significant impact of physical therapy activities conducted on a given patient, and the level of physician activities on a given patient on explaining the variance in LOS. Positive unstandardized $\beta$ values for provider dose indicates that, controlling for other factors, increasing the provider dose leads to an increase in LOS for all provider type and related activities. Results showed stronger physician dose effect and repository therapy dose effect on LOS. Previous studies on nursing showed positive effects of nursing staffing on LOS at the aggregate level and our study captured the effect of direct nursing care and was able to confirm the association between of many different nurses providing the care at the initial stages on increasing the LOS. One of the other important features of this study was that we analyzed the overall shared variability by combining the effects of all interdisciplinary provider types and related activities.

Findings of this study can be useful to hospital administrators when making daily work assignments. As higher provider workload driving longer LOS, for example, with nurses’ work assignments, hospital administrators may consider a balance between the severity of patients and number of patients. Also, with results indicating higher provider dose at the initial days of care driving longer LOS, for example, they may consider assigning the same physical therapist, as scheduling allows, for handling the patient care activities during the post-surgery rehabilitation. Overall, hospital administrators can utilize the results found in this study to identify better provider assignment, workload allocation, and scheduling strategies. Finally, while the final model with socio-demographic and clinical factors, and three provider factors accounted only for 32% variance there is ample opportunity for researchers to include a set of comprehensive clinical characteristics as well as define and evaluate additional provider factors to further decompose the variance in LOS.

**Limitation and Future work**

While we leveraged comprehensive EHR activity data to create provider factors, their definitions were prospectively defined by the investigators and represent an initial conception of interdisciplinary care team activity. In this study, we have pre-selected a general and limited set of clinical variables, for future studies we have recognize the need for including variables that have been identified to be driving the variations in length of stay through evidence based practices for LSS patients. As we focus further on robust LOS analysis, identifying rich clinical factors and care practices (medications, procedures, diagnosis, complications, labs, functional measures, alerts, etc.) through high dimensional feature (variable) selection machine learning algorithms may also be appropriate. Moreover, in this study, we have mainly focused on provider activity related factors, but acknowledge the importance of defining clear and consistent provider factors while investigating the effects of provider type, specialty, experience and credentialing on the variance in LOS. Thus, we hope to extend this study by integrating these factors as well as provider work assignment data from other sources such as staffing and assignment systems, human resource
information systems, and other public provider data repositories like National Plan and Provider Enumeration System (NPPES), Nursys, and the National Database of Nursing Quality Indicators.

Conclusion
Our study shows the feasibility of generating richer representations of care team activity from the EHR and its potential value in decomposing the unexplained variance in outcomes (LOS) attributed to the direct patient care actives performed by interdisciplinary providers. The quantity, patterns and timing of direct patient care actives performed by interdisciplinary providers may someday provide actionable indicators of patient outcomes, including LOS. Additional research is needed to determine how these findings could be used to improve health care quality and delivery.

Acknowledgement
This work was supported by a Clinical and Transitional Science Award (CTSA) grant from National Center for Advancing Translational Sciences (NCATS) awarded to the University of Kansas for Frontiers: University of Kansas Clinical and Translational Science Institute. (#UL1TR002366) The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of Health or NCATS.

References


30. Wells N, Pasero C, McCaffery M. Improving the quality of care through pain assessment and management.


Achievability to Extract Specific Date Information for Cancer Research

Liwei Wang, MD, PhD1, Jason Wampfler¹, Angela Dispenzieri, MD1, Hua Xu, PhD², Ping Yang, MD, PhD³, Hongfang Liu, PhD¹

1Mayo Clinic, Rochester, MN, USA; 2The University of Texas Health Science Center at Houston, Houston, TX, USA; 3Mayo Clinic, Scottsdale, AZ, US

Abstract

Accurate identification of temporal information such as date is crucial for advancing cancer research which often requires precise date information associated with related cancer events. However, there is a gap for existing natural language processing (NLP) systems to identify dates for specific cancer research studies. Illustrated with two case studies, we investigated the feasibility, evaluated the performances and discussed the challenges of date information extraction for cancer research.

Introduction

Cancer is the second leading cause of mortality in the US and despite extensive research and rapid advancements made in understanding this complex and heterogeneous disease, further innovative research leveraging real-world patient data could spearhead major therapeutic development. Observational studies, classically used to identify risk factors and prognostic indicators, were shown to be effective in estimating treatment effects¹. Observational studies have the advantage of reduced cost, being timely and offering a broad range of observable patterns making them an appealing method for cancer research. However, in order to draw conclusive results, labor- and time-intensive conventional data collection methods are required for observational studies².

Over the last decade, Electronic Health Records (EHR) systems have been increasingly implemented at US hospitals. Huge amounts of longitudinal and detailed patient information, including lab tests, medications, disease status, and treatment outcome, have been accumulated and are available electronically. One of the challenges faced in conducting EHR-based cancer research is to extract information from clinical narratives. Cancer-related events can vary among different cancers. For example, “transplant” and “conditioning” before bone marrow transplant therapy are specific events for multiple myeloma. Traditional chart review has been labor intensive and expensive. With the advancement of cancer informatics research, natural language processing (NLP) techniques have been explored to extract cancer-related events with some success. In cancer research, accurate identification of temporal information associated with cancer-related events is crucial ³. For example, diagnosis dates and treatment start dates of cancers are both important for survival analysis ⁴,⁵, postsurgical death evaluation within 30 days ⁶ or evaluation of post treatment side effect ⁷. Pinpointing diagnosis dates of cancer at various stages are crucial for selecting therapy ⁸,⁹, therefore, it’s very important for personalized cancer therapy ¹⁰. Earlier diagnosis dates were also associated with more favorable outcomes among specific cancer types ¹¹. In addition, initial treatment date is an important factor for effective cancer management ¹² and better health outcomes ¹³.

Existing clinical NLP efforts focus on profiling time line instead of pinpointing exact dates associated with specific events for unstructured data. For example, the Sixth Informatics for Integrating Biology and the Bedside (i2b2) Natural Language Processing Challenge for Clinical Records ¹⁴ was to extract three types (before, after and overlap) of temporal relations between the given events and between events and temporal expressions with a focus on temporal reasoning in clinical narratives. The Clinical TempEval challenges in SemEval from 2015 to 2017 also addressed temporal information extraction and temporal relation tasks. The evaluation was conducted separately for time expression, event and temporal relations. Temporal relations (before, after, overlap, before_overlap), which is most related to the above issue of extracting dates associated with specific cancer events, were evaluated at two-level through 1) the most coarse level, i.e., relating events to the document creation time; 2) narrative container relations, i.e., the event occurred within a certain time frame, which is not specifically associated with a date. ¹⁵⁻¹⁸

Systems produced ¹⁹⁻²² based on the above tasks are capable to extract temporal relations only on document level between events and document creation time or between events and a certain time frame. Gaps exist in identifying date information for specific EHR-based cancer studies: 1) they mostly use documentation dates as anchor dates; 2) they are not capable to extract exact dates associated with specific events across clinical documents of the same patient; and 3) they have not been practically used and evaluated in real clinical context.
There are very few existing studies focusing on date extraction for observational studies. One study attempted to automatically generate an epidemiological line list for real-time monitoring and responses to emerging public health threats, disease onset date was identified with the accuracy of 0.01 to 0.37, outcome date with the accuracy between 0.36 to 0.66 using word embeddings. However, this was a population-level study, without emphasis on exact date extraction for precise research purpose. Ruud et al. used the SAS text mining tool (SAS Text Miner) to extract date, time, physician, and location information of follow-up appointment arrangements from 6,481 free-text dismissal records at Mayo Clinic. The 6,481 free-text dismissal records were manually reviewed by a health services analyst to determine whether the instructions contained follow-up appointment arrangements. Sensitivity of date extraction achieved 0.996 (0.994–0.998) and specificity achieved 0.842 (0.828–0.856). Nevertheless, the study only focused on dismissal records of in-patient encounters.

In this study, we investigated the date extraction task using two case studies: one from malignant solid tumors (lung cancer) and one from malignant liquid tumors (multiple myeloma), this study investigated the feasibility, evaluated the performances and discussed the challenges of date information extraction for EHR-based cancer research, which to our knowledge has not previously been studied.

**Methods**

Figure 1 shows the overall study design for the two case studies, i.e., lung cancer and multiple myeloma. We developed a rule-based information extraction system using the open source clinical NLP pipeline MedTagger as the platform. The system consists of two steps: (1) identifying events: we first identified sentences with events from electronic health records (EHR). Negated events were removed. (2) extracting dates associated with events: we then extracted dates within or around these sentences and linked the dates with extracted events. Evaluation was conducted against the reference standards after normalizing dates with MedTime and heuristic rules. The following sections will elaborate each part in detail.

![Figure 1. Overall study design.](image)

**Date sets**

For lung cancer, an existing lung cancer cohort containing 4,110 patients definitively diagnosed with primary lung cancer from 1997 to 2016 was used for this study. Previously human abstractor manually reviewed charts to extract diagnosis dates for each patient. In this study, clinical narratives from various data sources including pathology reports and clinical notes were retrieved from Mayo Clinic clinical data warehouse (CDW). We randomly selected 499 lung
cancer patients and associated clinical narratives as the training set for deriving language expression patterns. The remaining 3,611 patients were used as the test set to evaluate the NLP system.

For multiple myeloma, we randomly selected 318 patients with a treatment date after year 2000 from an existing multiple myeloma database consisting of 11,656 patients collected over the past 50 years. Human abstractors have manually reviewed charts to extract treatment start dates for each patient. Similar to lung cancer, we retrieved the associated clinical narratives from Mayo Clinic CDW. Among the 318 patients, 185 patients have information on chemotherapy and 133 patients have information on transplant. For chemotherapy, 93 patients were used for development and 92 were used for testing, for bone marrow transplant 70 were used for development and 63 were used for testing. Transplant data elements include transplant type, transplant mobilization and transplant conditioning. However, there were gaps in the data abstraction as for the patients who underwent stem cell transplant, not all patients had their mobilization and conditioning information extracted. In the training set, only 17 out of 70 patients had mobilization fields completed and 26 had conditioning fields completed. In the testing set, the respective numbers were 26 out of 63 and 41 of 63.

**NLP method**

For lung cancer, the event to be extracted was lung cancer diagnosis. We first identify the sentences with mentions of lung cancer diagnosis by developing a custom dictionary of terms that described the histological cell types of lung cancer, examinations of lung, symptoms of lung cancer and positive malignancy of lung tumor. This dictionary was developed using the training data set. For pathology reports, histological cell types of lung cancer and positive malignancy of lung tumor were used to identify lung cancer in the diagnosis section. For clinical notes, the whole dictionary was used associated with the mention of “diagnosis” to identify the sentence with lung cancer diagnosis. After extracting lung cancer diagnosis, we then extracted dates associated with this event from both pathology reports and clinical notes.

To extract dates associated with lung cancer diagnosis in pathology reports. A two-step strategy was developed:

1. If the lung cancer diagnosis was identified in pathology reports without the section of “Path Review of Outside Specimen”, then extract the dates of the pathology reports. Outside specimen refers to the specimen prepared by other institutions instead of Mayo Clinic.
2. If there is a section of “Path Review of Outside Specimen” in the pathology report, and the lung cancer diagnosis was identified from the outside specimen, then extract the date associated with the outside specimen.

To extract dates associated with lung cancer diagnosis in clinical notes. A three-step strategy was developed:

1. If a date was appearing in the same sentence with lung cancer diagnosis, then extract the date.
2. If there is no date appearing in the same sentence with lung cancer diagnosis, but there is a date appearing in the sentences surrounding that sentence with lung cancer diagnosis, then extract the date.
3. If there is no date appearing in the same sentence or surrounding sentences with lung cancer diagnosis, then use the clinical date as the date of lung cancer diagnosis.

For multiple myeloma, events to be extracted included chemotherapy and bone marrow transplant. Bone marrow transplant has three data elements, i.e., conditioning, mobilization and transplant type, each of them has multiple values (Table 1). To identify sentences with mentions of treatments, we also employed the dictionary lookup approach by constructing a dictionary consisting of chemotherapy drugs and transplant data element values. Additional keywords found during the training process were added into the dictionary.

**Table 1.** Values of three transplant data elements.

<table>
<thead>
<tr>
<th>Values of Transplant type</th>
<th>Values of Conditioning</th>
<th>Values of Mobilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allogenic</td>
<td>1=Mel 200mg/m2</td>
<td>1=Cyclophosphamide</td>
</tr>
<tr>
<td>Autologous</td>
<td>2=Mel 140mg/m2</td>
<td>2=G-CSF</td>
</tr>
<tr>
<td></td>
<td>3=Mel oth/unk dose</td>
<td>3=G-CSF+Plerixafor</td>
</tr>
<tr>
<td></td>
<td>4=Mel-TBI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5=Cyclophosphamide/TBI</td>
<td></td>
</tr>
</tbody>
</table>
After identifying the values of chemotherapy and transplant data elements using the dictionary, a four-step strategy was developed to extract dates:

1. If a date was appearing in the same sentence with values of chemotherapy or transplant data elements, then extract the date.
2. If there is no date appearing in the same sentence with values of chemotherapy or transplant data elements, but there is a date inference in the sentence such as “currently receiving bortezomib, day +7”, then we use the clinical note date as the anchor date to infer the actual date.
3. If there is no date or date inference appearing in the same sentence with chemotherapy or transplant data elements, but there is a date appearing in the sentences surrounding that sentence, then extract the date. In our study, we first label the anchor sentences with any mentions of values of chemotherapy or transplant data elements. Then one sentence before the anchor sentence and two following sentences after the anchor sentence were extracted.
4. If there is no date appearing in the same sentence or surrounding sentences with values of chemotherapy or transplant data elements, then use the clinical note date as the anchor date of values of chemotherapy or transplant data elements.

After the above steps, the extracted date information can be linked to chemotherapy as conditioning and mobilization are pre-transplant procedures. Therefore, after extracting all dates associated with the values of the three transplant data elements, dates of transplant type were used as the anchor dates to identify dates associated with mobilization and conditioning. Conditioning usually was done two days before transplant, we used a time window of 4 days to identify date of conditioning associated with transplant type. Mobilization may be done as early as more than one month before transplant, we used a time window of 80 days to identify date of mobilization with transplant type.

Date Normalization

Dates extracted from pathology reports are all in the formats like “4/10/12” or “2012-04-10”. In contrast, dates extracted from clinical notes could be expressed in different formats, for example, “6/12/14”, “05-02-2016”, “Jan 7th, 2013”, “June 2015”, “April 17th”, or even in unprecise way such as “the end of March”, “early July”, “late 2012”, etc. Therefore, it is necessary to normalize the extracted dates before evaluating against the reference standards which are in the format of “2013-10-28”. A strategy of two-level date normalization was utilized. For the extracted dates with “yy-mm-dd”, “mm-dd”, “mm” or “yy-mm”, MedTime was used to normalize the dates. For the unprecise time, a set of heuristic rules were developed to map them to exact dates.

MedTime is the open source NLP pipeline for clinical temporal information extraction. It uses clinical note dates as anchor dates and can be used to normalize the extracted dates. Input dates with year, month and day in various formats as mentioned above will be normalized to the format of “YYYY-MM-DD”. If the input date has only information of year and month, it will be normalized to the day of “15”. For example, “June 2015” will be normalized to “2015-06-15”. If the input date has only information of month and day, the clinical note year will be used as the diagnosis year. For example, if “April 17th” appears in the clinical note recorded on “2014-02-14”, it will be normalized to “2014-04-17”.

The heuristic rules were primarily used to normalize obscure time of year, season and month. The following rules were used to normalize obscure time of year. Using such heuristics, a substitute date can be generated for all cancer events, from aggressive cancer types to multiple treatments that could occur over a short period of time, to prevent any missing data in data analysis.

- Exact mention of year was normalized to the middle date of the year, July 1, e.g., 1998 was normalized to July 1 1998.
• Early year: Feb 15 of the year, e.g., Early 2000 was normalized to Feb 15 2000.
• Beginning of year: January 15 of the year, e.g., Beginning of 2002 was normalized to January 15 2002.
• Middle of year: July 1 of the year, e.g., Middle of 2004 was normalized to July 1 of 2004.
• Late year: November 15, e.g., late 2006 was normalized to November 15 2006.
• End of year: December 15, e.g., end of 2008 was normalized to December 15 2008.

Four seasons were defined based on a three-month interval with spring starts from March and ends after May considering the weather in Rochester MN. The following rules were used to normalize obscure time of season.
• Exact mention of seasons was normalized to the middle date of the season, e.g., spring 2012 was normalized to April 15 2012.
• Taking “winter” as an example of obscure mention of seasons:
  a. Early winter: Dec 31, e.g., early winter 2001 was normalized to Dec 31 2001.
  b. End of winter: Feb 15, e.g., end of winter 2003 was normalized to Feb 15 2003.
  c. Late winter: Feb 1, e.g., late winter 2005 was normalized to Feb 1 2005.
  d. Middle of winter: Jan 15, e.g., middle of winter 2007 was normalized to Jan 15 2007.

The following rules were used to normalize obscure time of month.
• Exact mention of months was normalized to the middle date of the month, e.g., January 2011 was normalized to January 15 2011.
• Taking “June” as an example of obscure mention of months:
  a. Early or beginning of June: June 5, e.g., Early June 2012 was normalized to June 5 2012.
  b. Middle of June: June 15, e.g., Middle of June 2013 was normalized to June 15 2013.
  c. Late or end of month: June 25, e.g., Late June 2014 was normalized to June 25 2014.

Evaluation
Evaluation was conducted based on event value and date match within a time window of 6 days or 30 days. For lung cancer, it was counted as a match if extracted lung cancer diagnosis and date match with the reference standard within a time window of 6 days or 30 days. We randomly selected 20 from non-matched patients within 30 days and analyzed the reasons. For multiple myeloma, it was counted as a match if extracted values of chemotherapy, transplant type, transplant mobilization and transplant conditioning match with the reference standard within a time window of 6 days or 30 days. All un-matched cases within 30 days were reviewed for error analysis.

Results
Among the whole lung cancer cohort of 4,110 patients, 4,034 have 14,890 pathology reports, with an average of 3.7 notes per patient. 4,051 patients have 173,070 clinical notes, with an average of 42.7 notes per patient. Table 2 shows the evaluation results of diagnosis date extraction from pathology reports and clinical notes for the 3,611 testing patients. From the results matched within 6 days, diagnosis dates of 1,228 (34.0%) patients came from pathology reports done at Mayo, additional 1,777 (49.2%) came from the pathology reports with outside specimen reviewed at Mayo Clinic, combining clinical notes, 3,206 could be found achieving 88.8%. From the results matched within 30 days, diagnosis dates of 1,627 (45.1%) patients came from pathology reports done at Mayo, additional 1,638 (45.3%) came from the pathology reports with outside specimen reviewed at Mayo Clinic, combining clinical notes, 3,414 could be found achieving 94.5%.

Table 2. Evaluation results of diagnosis date extraction. Path (Mayo) refers to pathology report done at Mayo, Path (Outside) refers to pathology reports with outside specimen reviewed at Mayo.

<table>
<thead>
<tr>
<th>Match range</th>
<th>Data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Path (Mayo)</td>
</tr>
<tr>
<td>6 days</td>
<td>1228, 34.0%</td>
</tr>
<tr>
<td>30 days</td>
<td>1627, 45.1%</td>
</tr>
</tbody>
</table>

We analyzed the distribution of the missing diagnosis dates of 197 patients across different years. Figure 2 shows the time distribution of the missing diagnosis dates. Further analysis revealed that 13 patients have no clinical notes.
available. We randomly selected 20 patients in the remaining 184 patients and analyzed the reasons. Table 3 shows the error analysis results.

Table 3. Results of error analysis for lung cancer diagnosis date.

<table>
<thead>
<tr>
<th>Types of Error</th>
<th>Subtypes of Error</th>
<th>Number, %</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outside material</td>
<td>None</td>
<td>9, 45%</td>
<td>20</td>
</tr>
<tr>
<td>Human errors</td>
<td>None</td>
<td>5, 25%</td>
<td></td>
</tr>
<tr>
<td>Dictionary insufficiency</td>
<td>None</td>
<td>1, 5%</td>
<td></td>
</tr>
<tr>
<td>NLP rules</td>
<td>Missing inference for lung cancer diagnosis</td>
<td>3, 15%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing inference for dates</td>
<td>1, 5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date ambiguity</td>
<td>1, 5%</td>
<td></td>
</tr>
</tbody>
</table>

The most common errors are due to outside materials (45%) and human errors (25%) (Table 3). When patients came from another hospital or country, the information of diagnosis date was often not complete in clinical notes and pathology reports. Human error may primarily result from typing process where a very near keyboard number was mistakenly typed in. For example, diagnosis date of “2013-01-22” maybe typed as “2011-01-22”. Dictionary insufficiency accounted for 5%. For those errors from NLP rules, the primary challenge came from the inference for lung cancer diagnosis, accounting for 15%. This mainly came from patients who were referred from other institutions and experience a long journey for lung cancer diagnosis. We only integrated date inference piece into the NLP tool without “year” inference, the case (5%) missing from NLP needed to use “year” inference, i.e., “Since her last evaluation one year ago, she was diagnosed as having a non-small lung cancer”. Date ambiguity happens when two dates appeared in one sentence, such as clinical note date and diagnosis date, it’s hard to be certain which date should be used.

For multiple myeloma, 312 out of 318 patients have 44,976 clinical notes, with an average of 144 notes per patient. Table 4 shows the evaluation results of treatment dates extracted from clinical notes for the testing patients. 60 (65.2%) out of 92 chemotherapy treatment start dates matched with the reference standard within 6 days, and additional 11 (12%) were matched within 30 days. For transplant type, only 1 more case was found when using 30 days match range compared with 6 days match range. For mobilization, 2 more cases were found when using 30 days match range compared with 6 days match range. For conditioning, no more cases were found when using 30 days match range compared with 6 days match range.

Table 4. Evaluation results of treatment start date extraction.

<table>
<thead>
<tr>
<th>Match range</th>
<th>Chemotherapy (92)</th>
<th>Transplant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type (63)</td>
<td>Mobilization (26)</td>
</tr>
<tr>
<td></td>
<td>Conditioning (41)</td>
<td></td>
</tr>
<tr>
<td>6 days</td>
<td>60, 65.2%</td>
<td>49, 77.8%</td>
</tr>
<tr>
<td></td>
<td>17, 65.4%</td>
<td>24, 58.5%</td>
</tr>
</tbody>
</table>
Table 5 shows the error analysis for chemotherapy treatment date. Around half of 21 unmatched cases were due to human errors and outside materials (52.4%). In the remaining unmatched cases, the most challenge part came from lacking semantic inference for dates in NLP rules (33.3%). These cases were mostly referral patients who had a history of chemotherapy treatment in other institutions, and the clinical notes at Mayo Clinic only recorded the obscure time. For example, “She initially took thalidomide. She took treatment for about one to two years and then was observed until 2012”; or “He has been on ibrutinib for nearly three years”. When the chemotherapy in the reference standard was “Other” without clear definition of drug names, it is also challenging to extract corresponding chemotherapies. The last error came from the unprecise date mentioned in clinical note where only “year” was recorded. Through our heuristic rules, the year was normalized to “07-01” of the year, but the reference standard used “01-01”.

Table 5. Results of error analysis-chemotherapy start date.

<table>
<thead>
<tr>
<th>Types of Error</th>
<th>Subtypes of Error</th>
<th>Number, %</th>
<th>Total No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human errors</td>
<td>None</td>
<td>6, 28.6%</td>
<td>21</td>
</tr>
<tr>
<td>Outside material</td>
<td>None</td>
<td>5, 23.8%</td>
<td></td>
</tr>
<tr>
<td>Dictionary insufficiency</td>
<td>None</td>
<td>2, 9.5%</td>
<td></td>
</tr>
<tr>
<td>NLP rules</td>
<td>lacking semantic inference for dates</td>
<td>7, 33.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heuristic rules for unprecise dates</td>
<td>1, 4.8%</td>
<td></td>
</tr>
</tbody>
</table>

Table 6 shows the error analysis for dates of transplant data elements. Outside material accounts for 75% for transplant type. Human error accounts for 71.4% for mobilization and 76.5% for conditioning. The “other” value was not clearly defined in dictionary. Thus, it was hard to extract. Since date mentions were very diverse, one was missed in the NLP rules.

Table 6. Results of error analysis-transplant date.

<table>
<thead>
<tr>
<th>Data elements</th>
<th>Types of Error</th>
<th>Number, %</th>
<th>Total No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transplant type</td>
<td>Outside material</td>
<td>9, 75%</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>No notes</td>
<td>3, 25%</td>
<td></td>
</tr>
<tr>
<td>Mobilization for transplant</td>
<td>Human error</td>
<td>5, 71.4%</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>No notes</td>
<td>2, 28.6%</td>
<td></td>
</tr>
<tr>
<td>Conditioning for transplant</td>
<td>Human error</td>
<td>13, 76.5%</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Dictionary insufficiency</td>
<td>2, 11.8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Outside</td>
<td>1, 5.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NLP rule</td>
<td>1, 5.9%</td>
<td></td>
</tr>
</tbody>
</table>

Discussion

Temporal extraction systems have not been transportable from one condition (e.g., colon cancer) to another (e.g., brain cancer) as shown in the SemEval challenge. When training and testing on only notes from colon cancer patients, the top system achieved F1s of 0.76 for document time relations in 2016 with an increase of 0.058 compared with 2015, and 0.48 for narrative containers with an increase of 0.22 compared with 2015. When training on notes from colon cancer patients and testing on notes from brain cancer patients, the best F1 achieved 0.50 in linking events to document creation time and above 0.30 F1 for linking events to their narrative containers, with a 20+ point drops in performance. Different patient cohorts and linguistic patterns behind different cancers are the underlying reasons for the results.

In the two case studies used in this research, lung cancer and multiple myeloma, all clinical notes were obtained from Mayo Clinic CDW which recorded not only events happened at recording time, but also recorded historical events with associated dates, if any. But different features of date expressions can be found due to distinct differences of various events in the two different cancers. Most multiple myeloma patients have clear records with treatment start dates. Since almost half of lung cancer patients in the cohort were referred from other institutions, learnt from the
results (Table 2), in the clinical notes the history of lung cancer may be recorded very unclearly with more obscure time used than the multiple myeloma case.

Therefore, extracting dates associated with various cancer events required different event extraction dictionaries, different date extraction strategies as well as different approaches for linking dates with extracted events. Our study demonstrated that it is feasible to extract exact dates associated with specific cancer events for EHR-based cancer research. However, some challenges exist as illustrated in the following.

To extract precise dates associated with cancer events, it’s crucial to assert events first followed by temporal information extraction. However, it’s hard to accurately extract events as the events scatter in different data sources and there are no definite event mentions, as shown in the lung cancer case study. When using different data sources, different NLP strategies were necessary to extract events. Pathology reports record lung cancer diagnosis mainly through specific histologic cell types or assertion of positive malignancy in lung related biopsy sites. While in clinical notes lung cancer diagnosis is often represented by other miscellaneous terms, e.g., from lung examinations and symptoms of lung cancer in addition to specific histologic cell types or assertion of positive malignancy. This needs a large amount of time for developing dictionary during the training process if there was no one available. Second, the diagnosis of lung cancer may be a long journey and lacking definite diagnosis mentions was common. When the initial examinations provided uncertain results, the suspicious results would just remain undecided in one facility until it was ascertained in another facility. During the process, the mention of date and diagnosis are often existing in sentences far away from each other. To capture this complex situation, the date and diagnosis across multiple sentences were extracted using NLP. However, it’s hard to accurately extract diagnosis and date when too many sentences were involved, especially when there was no definite diagnosis mentions and there was a need to infer.

Since a lot of patients have been referred to Mayo Clinic for further treatment from other institutions, recording of exact diagnosis dates are not very clear in some cases. For example, lung cancer diagnosis date may be mentioned as “early winter 2000”. Given the long history of cancer registry databases, inconsistency of date normalization may exist among human abstractors to normalize the obscure time to specific dates. The gap may also exist between the heuristic rules we developed and the rules human abstractors used. In addition, different physicians have different styles for recording notes and there were no unified rules for recording dates. For example, it’s hard to normalize the chemotherapy start time recorded as “05/12”, because it’s hard to determine if “05” was month or year, and if “12” was month or “day”.

Linking dates to the associated events lies in the fact that events may be dependent to each other. As shown in the multiple myeloma case study, both transplant mobilization and conditioning affiliated to transplant type. Extracted mobilization and conditioning dates were linked to transplant type dates using inference rules, which may result in errors. Linking mobilization and conditioning values to associated dates was hard because 1) some patients may have several transplants within short time intervals, 2) different mobilization and conditioning values may appear repeatedly in clinical notes. This poses challenges not only for NLP but also human. In fact, as shown in evaluations we found human made many errors when linking mobilization and conditioning values to associated dates which NLP had successfully avoided.

In addition, events may have multiple related dates and it’s hard to choose the optimal date, as shown in both case studies. For lung cancer, multiple dates associated with definite lung cancer diagnosis existed in pathology reports. For example, one patient may have the same lung cancer diagnosis from different pathology reports on different dates using different specimens. For multiple myeloma, some patients may have as many as more than 20 times of chemotherapies, some of which were within very short time intervals. Thus, NLP techniques can accelerate the data abstraction process but not completely replace human efforts.

Conclusion

It is feasible to extract exact dates associated with specific cancer events for EHR-based cancer research using automatic NLP method to facilitate cancer registry curation. However, data extraction in various cancer registry databases has been ad hoc for specific research purpose. To gear with the demands for research-oriented information collection, developing NLP algorithms needs to consider the differences among various cancers in event extraction.
dictionaries, date extraction strategies as well as approaches for linking dates with extracted events. We plan to extend our investigation to include more cancer patients and more cancer types.

Acknowledgement

This work was supported by National Cancer Institute 1U24CA194215-01A and U01TR02062. We acknowledge Dr. Han Liu, Dr. Yanqi He, Dr. Lin Du, Dr. Yung-Hung Luo and Kathryn M Johanns for data annotation.

References


Towards augmenting structured EHR data: a comparison of manual chart review and patient self-report

Nicole G. Weiskopf, PhD, Aaron M. Cohen, MD, MS, Joely Hannan, BS, Thad Jarmon, MBA, David A. Dorr, MD, MS
Department of Medical Informatics and Clinical Epidemiology, OHSU, Portland, OR

Abstract
Structured electronic health record (EHR) data are often used for quality measurement and improvement, clinical research, and other secondary uses. These data, however, are known to suffer from quality problems. There may be value in augmenting structured EHR data to improve data quality, thereby improving the reliability and validity of the conclusions drawn from those data. Focusing on five diagnoses related to cardiovascular care, this paper considers the added value of two alternative data sources: manual chart abstraction and patient self-report. We assess the overall agreement between structured EHR problem list data, abstracted EHR data, and patient self-report; and explore possible causes of disagreement between those sources. Our findings suggest that both chart abstraction and patient self-report contain significantly more diagnoses than the problem list, but that the information they capture is different. Methods for collecting and validating self-reported medical data require further consideration and exploration.

Introduction
Electronic health record (EHR) data serve a number of important roles beyond their primary purposes related to patient care, including clinical research, algorithm development and machine learning, epidemiologic or population-based studies, or in measuring the quality and safety of care. In the age of the Quality Payment Program and other value-based care initiatives, one of the most important uses is to calculate and report clinical quality measures (CQMs). CQMs are written with varying levels of formalization and adherence to standards. At one end of the spectrum, measures included in the Centers for Medicaid and Medicare Services quality reporting program are specified as electronic CQMs (eCQMs). These eCQMs are currently written in the Health Level 7 Health Quality Measure Format and employ Quality Data Model logic; in the future eCQMs will be written in Clinical Quality Language, also maintained by Health Level 7. The clinical concepts required to calculate CQMs are encoded in official value sets using a variety of medical standards and terminologies. The value sets are available at the Value Set Authority Center, which is maintained by the National Library of Medicine. CQMs that are not fully specified range in formality from plain language to carefully specified logic. These CQMs also rely on the availability of certain key clinical concepts, and therefore list specific concept codes for medications, diagnoses, and procedures.

Regardless of whether specific concept codes are provided in measure specifications, the most efficient way to calculate CQMs, electronic or otherwise, is through the reuse of structured EHR data. Structured data may be used in native EHR quality calculation engines, or extracted from the EHR via automated processes, enabling consistent and continuous (or at least regular) calculation. Unfortunately, there are limitations to structured EHR data, including incompleteness, inaccuracy, and non-current data. Chan et al.’s 2010 literature review found that the quality of clinical concepts needed for quality measure calculation was highly variable and often poor. How best to assess, understand, and improve the quality and reliability of structured EHR data remains an open question. One limiting factor is the fact that traditional approaches to data quality assessment often rely on the availability of a gold standard, which is one of the most common methods of EHR data quality assessment as well. The problem with this approach is that it is very rare to have an actual gold standard available, meaning that most such efforts must rely upon comparison against other sources of information that are themselves likely to suffer from limited validity and reliability (e.g., paper medical records or direct observation of clinical encounters).

Rather than assume that one single source of medical information is the gold standard against which other sources can be compared, it may be best to approximate the truth by combining data from multiple sources. Structured data represent only a small portion of the data contained within the EHR, which also includes large quantities of narrative notes and other forms of non-structured data, such as images and scanned files. In fact, it has been shown that structured and unstructured EHR data do not always agree. Another alternative source of medical information is patients themselves. There is substantial literature on comparisons between self-report and the medical record, with highly variable findings that are impacted by clinical condition of interest, demographic factors, or data collection methods. For example, Vigen et al. measured agreement between self-reported diabetes and hypertension diagnoses in breast cancer survivors and clinical data extracted from Kaiser Permanente’s Virtual Data Warehouse, which was constructed from a combination of EHR data, outpatient pharmacy data, and
administrative claims data. They reported high levels of agreement for both diagnoses (diabetes $\kappa = 0.87$; hypertension $\kappa = 0.80$). In contrast, Wagaw et al. looked at the same two diagnoses and compared structured EHR data from two urban health clinics against patient self-report, and found only moderate agreement for both diagnoses (diabetes $\kappa = 0.76$; hypertension $\kappa = 0.69$). In the Wagaw study, the research was limited to data from two clinics, whereas the Kaiser Permanente clinical database contained medical records from multiple locations and sites of service, including pharmacy data. Another reason for the disparity in results are the methods of collecting self-report data. In the Wagaw study, self-reported data were collected using a self-administered questionnaire, while Vigen et al. collected patient self-report via in-person interviews.

While the limitations of EHR data quality and the variable level of agreement between EHR data and patient self-report indicate that there is no single gold standard when it comes to medical information, the discrepancies between these data sources provide an opportunity to bring these sources together to create a more comprehensive, timely, and accurate picture of patients' medical status and histories than what we glean from purely structured data. Therefore, to develop a deeper understanding of the limitations of structured EHR data and explore options for augmenting these data, we compared the presence of five diagnoses related to cardiovascular health across three data sources: structured problem list data, manual chart abstraction data, patient self-report data collected via a survey. Our goal was to characterize and quantify agreement between these data sources and to begin to identify and understand the underlying factors that may lead to disagreement between these sources.

**Methods**

We measured agreement between three data sources on a set of five diagnoses related to cardiovascular health. The three data sources were: 1) the structured problem list queried from the OHSU instance of the Epic relational database; 2) manually abstracted diagnosis information from the complete medical chart available through the OHSU Epic EHR instance; and 3) patient self-report, collected via survey. Per eCQM specifications, encounter diagnoses – not problem list – are used for calculation; however, patient self-report requires a global response similar to the problem list (i.e., at the patient level, not encounter level). The trade-off in the temporality and sensitivity of encounter diagnoses versus the precision (positive predictive value) of the problem list was thought to favor the use of the problem list for the purposes of this comparison. The diagnoses included were diabetes, hypertension, heart failure, myocardial infarction, and stroke. Level of agreement was determined via a combination of descriptive statistics, paired tests of proportion, and Fleiss’ kappa. Error analysis was conducted via manual chart review to uncover causes of disagreement and implications for augmentation of the structured problem list data.

**Subjects, Recruitment, and Institution**

The population of interest was defined as patients between the ages of 18 and 89, with English listed as their preferred language, having had at least one outpatient visit with the Oregon Health & Science University (OHSU) cardiovascular department within the previous calendar year as of 2/12/2018. OHSU is an academic medical center in Portland, Oregon, that includes two hospitals, with a total of 572 beds, and a number of ambulatory care clinics. From this population, a random sample of 4,000 patients was selected as potential participants. Batches of invitations to participate in a brief survey regarding their medical care and history were sent to these patients via email until at least 300 participants initiated the survey. The respondents comprised the study sample for all three data collection arms. Chart review was not performed on those records labeled as “restricted” within the OHSU Epic instance. This study was approved by the OHSU Institutional Review Board (#00017632).

**Data and Concept Extraction**

This study is part of a larger project to examine the impact of data quality on clinical quality measure (CQM) reliability. The CQMs used in this project cover a combination of four general cardiology measures and a set of specialty heart failure measures. These CQMs require a number of clinical concepts in order to calculate denominator inclusion and numerator adherence, including several diagnoses related to cardiovascular health.

*Problem List Data:* Diagnoses were queried from the problem list in Epic via the Integrated Care Coordination Information System, a population management system. Relevant active diagnoses were identified using ICD-10-CM codes from diagnosis value sets downloaded from the Value Set Authority Center maintained by the National Library of Medicine (diabetes mellitus: 2.16.840.1.113883.3.464.1003.103.12.1001, hypertension: 2.16.840.1.113883.3.464.1003.104.12.1011, myocardial infarction: 2.16.840.1.113883.3.526.3.403, stroke: 2.16.840.1.113762.1.4.1047.44, and heart failure: 2.16.840.1.113883.3.526.3.376). Each diagnosis was coded as either present or absent for each participant and had to be present at or before the time the survey was sent out.

*Patient Self-Report Data:* While self-reported medical history is often collected in the medical setting, we were unable to find a validated instrument that addressed the combination of concepts required for this research.
Therefore, a REDCap survey was designed to assess relevant diagnoses, medication usage, tobacco use and therapy, laboratory results, and heart failure status and treatment (the analyses described in this manuscript focus only on the diagnoses).\textsuperscript{11} The survey was designed with input from and review by cardiology providers at OHSU, and sought to balance depth and breadth of information collected with instrument brevity and required health literacy. Diabetes mellitus, myocardial infarction, stroke, and hypertension were assessed with forced binary checkboxes indicating presence or absence of the diagnosis. The diagnosis of heart failure, which the cardiology providers decided was more complex, was described in greater detail and used Likert-like response options: definitely not, maybe not, unsure, maybe yes, and definitely yes. For the purposes of the analyses described in this paper, the responses of maybe yes and definitely yes were considered to be affirmative, and any other response was considered negative.

**Chart Review Data:** Diagnoses were extracted from the OHSU Epic instance by one of three trained chart reviewers. For each of the five diagnoses, an annotation guide was created that included a list of diagnosis names, abbreviations, and terms based on the official value sets described above and review of common medical terminology. The reviewers used these term lists to search for structured and unstructured references to those diagnoses anywhere in the patient chart, which were then reviewed in greater detail. When relevant results were not found via the search tool, the reviewers were instructed to also review other information sources within the chart that could not be searched in an automated fashion; these other sources included PDFs of reports and letters from providers and linked records from partner institutions and providers available from health information exchange. Chart reviewers were told not to make interpretations based upon evidence supporting a diagnosis (e.g., laboratory results or medication orders), but rather to look for explicit diagnoses documented by providers. As with the problem list data, each diagnosis was coded as either present or absent for each participant.

**Analysis**

The overall goal of the analysis was to determine the degree of overlap and difference between the three data sources, as well as to understand the reasons why there was disagreement. Outcomes included similarity of proportions across the sources, agreement between the sources, and error analysis for the discrepancies. To begin, the presence of each of the five diagnoses in each of the three data sources was determined at the participant-level and then assigned to one of the seven possible categories of agreement, as summarized in Figure 1.

**Statistical tests:** For each diagnosis, Cochran’s Q test was used to test whether the proportion of diagnoses present was the same for each data source. For diagnoses where there was determined to be a significant difference in proportion, post-hoc analyses comparing pairs of data sources were conducted using McNemar’s test. Bonferroni corrections were applied to the original alpha of 0.05, resulting in an alpha of 0.01 for the Cochran’s Q tests and 0.003 for the McNemar tests. Agreement between the three sources for each diagnosis was determined using Fleiss’ kappa (\(\kappa\)). Proportional Venn diagrams were created for each diagnosis using the software program eulerAPE.\textsuperscript{12}

**Error analysis:** Error analysis was conducted on a sample of cases where there was disagreement between the data sources. Where present, two cases were selected for review for each agreement category and each diagnosis. For each case, an attempt was made to uncover the source of disagreement for that diagnosis through further review of the chart, value sets, and system processes. The chart-based error analysis was conducted by one of the trained chart reviewers (JH), and was reviewed by the principal investigator (NGW) to ensure consistency of findings. Where a cause of the disagreement could not be determined via chart review, the principal investigator reviewed the relevant queries, diagnosis coding processes, and value sets in an effort to uncover disagreement causes related to the backend systems. Once the disagreement sources were identified, a deductive approach was used to iteratively review and categorize the sources in order to identify a set of key factors resulting in the disagreement.

**Results**

We sent survey invitations to 1,700 eligible patients, 307 (18\%) initiated the survey by clicking yes on the consent page and completed the section of the survey relating to medical history. Of those 307 respondents, 32

![Figure 1. The three data sources (problem list, chart review, and self-report via survey) represented as sets in a Venn diagram. Overlapping segments indicate participants with the same diagnosis found in more than one data source.](image-url)
had restricted charts in our Epic instance, and were therefore not included in our analyses, for a final sample of 275 participants.

Mean participant age was 60.1 years, with a standard deviation of 15.3 years. Participants were uniformly distributed according to sex (51.6% female), and disproportionately white (94.2%) and non-Hispanic (97.1%). Participants had a mean of 4.9 ± 4.0 unique diagnoses on their problem lists and 10.5 ± 6.0 unique medication classes on their medication lists. See Table 1 for details.

Overall, 76% of the participants had at least one of the five diagnoses according to at least one of the three data sources, for a total of 380 distinct diagnoses. There was complete agreement for 182 of these cases (48%). Hypertension was the most commonly-reported diagnosis across all three data sources, and stroke the least common. More specifically, hypertension was reported in at least one of the three data sources for 63% of the participants, heart failure for 28%, diabetes for 20%, myocardial infarction for 16%, and stroke for 12%. The data source with the highest rate of endorsed diagnoses was the patient survey, in 70% of all participants reported at least one of the five diagnoses. Sixty-five percent of the participants had at least one of the diagnoses according to chart review, and 55% according to the structured problem list. The proportion of participants with each diagnosis by data source is summarized in Table 2.

Of the seven possible data source agreement categories, a diagnosis being present in all three sources was the most common category for each diagnosis is shown in Table 3. Of the seven possible data source agreement categories, a diagnosis being present only in the problem list or in the problem list and patient survey were the least common (2% and 1% respectively). Across all diagnoses combined, the most common category of data source agreement was when a diagnosis was present in all three sources (48%), and after that was when a diagnosis was only reported in the patient survey (22%). The complete breakdown of agreement category for each diagnosis is shown in Table 3.

Table 1. Demographic details of final sample of 275 participants.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>n</th>
<th>Mean ± SD</th>
<th>Female %</th>
<th>Male %</th>
<th>Hispanic %</th>
<th>Non-Hispanic %</th>
<th>Unknown or Declined %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>275</td>
<td>60.4 ± 15.3</td>
<td>51.6</td>
<td>48.4</td>
<td>10.5</td>
<td>89.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>142</td>
<td>51.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>133</td>
<td>48.4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian, Chinese</td>
<td>100</td>
<td>36.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>100</td>
<td>36.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean or Pacific Islander</td>
<td>50</td>
<td>18.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td>50</td>
<td>18.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White or Caucasian</td>
<td>250</td>
<td>92.2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown or Declined</td>
<td>25</td>
<td>9.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown or Declined</td>
<td>10</td>
<td>3.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique Medication Classes</td>
<td>105</td>
<td>3.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Percentage of cases reporting a diagnosis in: any of the three data sources, each data source separately, or all the data sources. Significant differences amongst the three data sources according to Cochran’s Q are indicated by asterisks in the second column (alpha = 0.01 with Bonferroni correction). The matched lowercase letters a and b indicate pairwise differences between two of the data sources based on McNemar tests (alpha = 0.003 with Bonferroni correction).

<table>
<thead>
<tr>
<th>Diagnoses</th>
<th>Any source</th>
<th>Problem list</th>
<th>Chart review</th>
<th>Self-report</th>
<th>All sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myocardial infarction</td>
<td>16% *</td>
<td>5% a,b</td>
<td>10% a</td>
<td>13% b</td>
<td>4%</td>
</tr>
<tr>
<td>Heart failure</td>
<td>28% *</td>
<td>15% a</td>
<td>19%</td>
<td>24% a</td>
<td>12%</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>20% *</td>
<td>14%</td>
<td>17%</td>
<td>19%</td>
<td>13%</td>
</tr>
<tr>
<td>Hypertension</td>
<td>63% *</td>
<td>43% a</td>
<td>51% a</td>
<td>52%</td>
<td>34%</td>
</tr>
<tr>
<td>Stroke</td>
<td>12% *</td>
<td>3% a,b</td>
<td>7% a</td>
<td>10% b</td>
<td>3%</td>
</tr>
</tbody>
</table>

Table 3. For each diagnosis reported in at least one of the three data sources, the total number of cases for which that diagnosis is present, and the percentage of those cases present in each category of agreement between the three sources. CR = chart review, PL = problem list, and SR = patient self-report via survey.

<table>
<thead>
<tr>
<th>Proportions of data source agreement categories for each diagnosis</th>
<th>total n</th>
<th>PL &amp; CR</th>
<th>PL &amp; SR</th>
<th>CR &amp; SR</th>
<th>PL only</th>
<th>CR only</th>
<th>SR only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myocardial infarction</td>
<td>45</td>
<td>24%</td>
<td>7%</td>
<td>0%</td>
<td>20%</td>
<td>9%</td>
<td>38%</td>
</tr>
<tr>
<td>Heart failure</td>
<td>76</td>
<td>45%</td>
<td>4%</td>
<td>3%</td>
<td>12%</td>
<td>3%</td>
<td>28%</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>54</td>
<td>69%</td>
<td>2%</td>
<td>0%</td>
<td>17%</td>
<td>2%</td>
<td>9%</td>
</tr>
<tr>
<td>Hypertension</td>
<td>172</td>
<td>54%</td>
<td>11%</td>
<td>1%</td>
<td>12%</td>
<td>2%</td>
<td>16%</td>
</tr>
<tr>
<td>Stroke</td>
<td>33</td>
<td>21%</td>
<td>0%</td>
<td>0%</td>
<td>21%</td>
<td>0%</td>
<td>15%</td>
</tr>
<tr>
<td>Any diagnosis</td>
<td>380</td>
<td>48%</td>
<td>7%</td>
<td>1%</td>
<td>14%</td>
<td>2%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Statistical results

Cochran’s Q tests showed statistically significant differences in proportions of positive cases between the three data sources for each of the five diagnoses after correcting for multiple hypotheses (adjusted p-value of 0.01). Post-hoc pairwise McNemar tests showed significant differences in proportion between the structured problem list and chart review for myocardial infarction, hypertension, and stroke; significant differences in proportion between the structured problem list and patient self-report for myocardial infarction, heart failure, and stroke; and no
significant differences in proportion between chart review and patient self-report (adjusted p-value of 0.003). Significant differences of proportions are indicated by asterisks for Cochran’s Q and lowercase matched letters for McNemar tests in Table 2.

Fleiss’ kappa indicated moderate agreement between the three data sources for myocardial infarction (κ = 0.52) and stroke (κ = 0.49), substantial agreement for hypertension (κ = 0.62) and heart failure (κ = 0.67), and the highest agreement for diabetes (κ = 0.85). Agreement is summarized in Figure 2 using Venn diagrams.\textsuperscript{12}

![Figure 2. Proportional Venn diagrams showing relative rates of diagnoses present in each of the three data sources. $\kappa$ indicates Fleiss’ kappa, CR = chart review, PL = problem list, SR = patient self-report via survey, and $n$ refers to the total number of cases in which the diagnosis was present in any of the three data sources.](image)

Error analysis: categories of factors causing disagreement

In each case of disagreement between the three data sources, the disagreement might be due to false positives in the source(s) in which the diagnosis was present or false negatives in the source(s) in which it was not present. We conducted error analysis on 57 (29\%) of the 198 cases where a diagnosis was present in only one or two of the data sources. Our analysis revealed five categories of factors that could lead to disagreement between the three sources via the presence of false negatives or false positives. These categories included research error introduced by our process, healthcare system factors, documentation factor, patient factors, and diagnosis factors. Problems with self-report were the most difficult to definitively identify and categorize, since, and in most cases these causes of disagreement could only be suspected, not definitely proven because the underlying truth could not be extrapolated from the information available in the chart. These factors are summarized in Table 4 and described in more detail below.

**Healthcare system factors:** We identified two situations related to the healthcare system that led to false negatives in the problem list data and chart review data. First, false negatives sometimes occurred in both the EHR-derived sources when primary care was provided outside of the OHSU system. A subset of these cases included patients receiving highly focused specialty care (e.g., vascular surgery) at OHSU. Both of these factors are related to the fact that healthcare is often delivered across multiple institutions and by multiple providers.

**Documentation factors:** We found two factors related to EHR documentation practices. False negatives and false positives in the problem list sometimes arose when the problem list was not consistently maintained and was therefore out-of-date, either because a resolved problem was not removed or because an active problem was not added (or was added after the measurement period concluded). The second factor in this category was the use of copy and paste, which sometimes led to false positives in the chart review data when an erroneous diagnoses, rather than being eliminated from the chart, were propagated forward in narrative notes over time.

**Patient factors:** Patients who are less engaged with their healthcare or who have lower health literacy might not fully grasp their own health status, leading to false negatives or false positives in self-report. The other patient-related factor, medical complexity, refers to the fact that patients with multiple active diagnoses and medical concerns may have incomplete medical charts (false negatives in the problem list or chart review data), and may also
have difficulty understanding the entirety of their own medical situation, leading to false negatives or false positives in the self-report. As noted above, these factors often could not be identified with certainty, but there were instances where low engagement was mentioned explicitly in the chart, or where medical complexity could be inferred.

**Diagnosis factors:** There were two types of factors identified in this category. First, the acuity of a diagnosis appeared to be associated with the presence of false negatives in the problem list. Acute and historic diagnoses (like a myocardial infarction or stroke experienced prior to the initiation of care at OHSU) were not always added to the problem list. Similarly, diagnoses that were chronic but well-controlled (like hypertension or diabetes controlled via lifestyle changes) were not always present on the problem list. The other factors in this category involved the existence of related diagnoses. First, diagnoses with semantic similarity sometimes led to false negatives or false positives in patient self-report, as when a patient conflated cardiac arrest and myocardial infarction. Second, there were cases when a diagnosis was initially expected, added to the problem list, then ruled out, but not deleted from the problem list or appropriately labeled in the chart as erroneous. For example, a patient might be admitted via the emergency department with a suspected myocardial infarction, only to have this diagnosis ruled out via testing.

**Research error:** This final category includes factors arising from our research process that led to false negatives or false positives in any of the three data sources. ETL errors sometimes led to resolved diagnoses being labeled as active in our problem list data (false positives) or missing certain diagnoses entirely (false negatives). Chart review errors led to false positives and false negatives in the chart review data. Value set problems occurred when a value set included codes and problem names that didn’t necessarily indicate the diagnosis of interest, which might lead to false positives or false negatives in the problem list data or the chart review data (e.g. Dressler’s syndrome is included in the myocardial infarction value set, but may have other causes). Finally, there were some known limitations involving the validity of the survey, especially with language that was either too clinical, leading to false negatives in the self-report data, or too broad, leading to false positives in the self-report data.

**Error analysis: categories of source disagreement**

**Problem list only:** Due almost exclusively to problem list false positives resulting from ETL errors (research error) or failure to maintain the problem list, leading to the inclusion of resolved problems.

**Chart review only:** Due either to false negatives in the problem list and self-report or false positives in the chart review. False negatives in the problem list data resulted from healthcare system factors and diagnosis-related factors, though false negatives in the self-report data were harder to diagnose. We also found some false positives in the chart review data stemming from documentation practices, specifically use of copy and paste, and chart review error (research error).

**Patient survey only:** Due either to false negatives in the two chart-derived data sources, or false positives in the self-report data. This form of disagreement was one of the most difficult to diagnose. Some of these cases were a combination of chart review error leading to false negatives in the chart data and diagnosis-related factors leading to false negatives in the problem list data. Self-report false positives may have occurred because of survey validity.

### Table 4. Error analysis revealed five categories of factors potentially leading to false negatives or false positives in the data sources.

<table>
<thead>
<tr>
<th>Category</th>
<th>Problem list</th>
<th>Chart review</th>
<th>Self-report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare system factors</td>
<td>FP</td>
<td>FN</td>
<td>FP</td>
</tr>
<tr>
<td>Primary care elsewhere</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Specialty care focus</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td><strong>Documentation factors</strong></td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Maintenance of problem list</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Use of copy and paste</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td><strong>Patient factors</strong></td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Medical complexity</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Health literacy</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Engagement with healthcare</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td><strong>Diagnosis factors</strong></td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Acuity</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Acute and historic</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Chronic and well-controlled</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td><strong>Related diagnoses</strong></td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Semantic similarity</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Suspected diagnoses</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td><strong>Research error</strong></td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>ETL error</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Chart review error</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Value set problems</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
<tr>
<td>Survey validity limitations</td>
<td>✕</td>
<td>✕</td>
<td>✕</td>
</tr>
</tbody>
</table>
limitations (research error), diagnosis-related factors like diagnoses with semantic similarity, or patient factors, like low health literacy leading to poor comprehension of the survey. In many cases the cause of disagreement could not be determined based on information available for error analysis in the chart. It is likely that in some of these cases patients were reporting information that was absent from the OHSU chart, likely due to healthcare system factors.

**Problem list and chart review, but not survey:** Due either to false negatives in self-report or false-positives in the two chart-derived data sources. Definitive causes could not be identified for most of these cases. Patient factors like low health literacy or low engagement, or diagnosis-related factors like the presence of well-controlled chronic diagnoses (therefore not the focus of their medical experience) leading to false negatives in survey data were a possible culprit. Alternatively, cases where a false positive was present in the chart review data, stemming from the use of copy and paste (documentation factor) or the presence of rule-out diagnoses (diagnosis-related factor), could have led to the presence of that same diagnosis on the problem list when it should have been deleted during problem list maintenance (documentation factor).

**Problem list and chart review, but not chart review:** This form of disagreement appeared to be largely, though not exclusively, related to temporality and time windowing problems leading to false positives in the problem list data and self-report data. There were cases when no-longer-active diagnoses were mistakenly captured from the problem list through our ETL process (research error) or were not removed from the problem list following resolution (documentation issues). Similarly, temporality was intentionally absent from the survey questions in order to improve readability, but at the expense of survey validity in some cases; diagnoses reported by patients were not always current, effectively leading to false positives in the survey (research error).

**Chart review and survey, but not problem list:** Generally the cases we reviewed in this category were due to false negatives in the problem list, though a combination of false positives in the chart review and self-report data would also be possible. Causes of problem list false negatives included diagnosis-related factors like acute historic events and chronic well-controlled diagnoses not being added to the problem list. There were also cases where a patient did not receive primary care at OHSU or received only very specific specialty care at OHSU (system factors) leading to diagnoses that were found in unstructured notes or other media, but which hadn’t been added to the problem list.

**Discussion**

This study explored the agreement between the structured EHR problem list, manual chart review, and patient self-report for five diagnoses commonly associated with cardiovascular health and treatment, in the service of understanding the potential value-added of the latter two data sources when combined with structured EHR problem list data. Tests of proportion showed significant differences between the three data sources for all five diagnoses, but no differences in proportion between the chart review and patient self-report data. Fleiss’ kappa coupled with proportional Venn diagrams revealed varying degrees of agreement between the data sources, with diabetes mellitus having the highest level of agreement, and stroke the lowest. Error analysis identified several different potential sources of disagreement between the three data sources. These potential causes of disagreement could be grouped into five broad categories: research error, healthcare system factors, documentation factors, diagnosis factors, and patient factors. While potential causes of disagreement could be identified in many cases, the degree of certainty about those causes varied, ranging from high confidence to strong suspicion to conjecture.

**Patterns of agreement between sources**

The diagnoses found in the problem list generally had high overlap with those found in via chart review, with the problem list diagnoses being largely a subset of those found via chart review, as shown in Table 3 and visualized in Figure 2. Exceptions occurred when chart review revealed that a diagnosis present on the problem list was erroneous or out-of-date, or when the ETL process used to collect the problem list data led to research errors. We also found that chart review captured a significantly larger proportion of diagnoses than the problem list for myocardial infarction, hypertension, and stroke. Error analysis showed that these differences in proportion might be due to diagnosis-related factors, including the fact that prior acute events and well-controlled chronic diagnoses were not always added to the problem list, even when they were noted in the unstructured medical history. Second, due to healthcare system factors, diagnoses that were not the focus of care at OHSU were less likely to be entered on the problem list. This aligns with previous work that found that primary care providers are responsible for the majority of problem list documentation.13

Chart review and patient self-report, by contrast, had similar proportions with no significant differences between the two sources for any diagnosis, but limited overlap, as shown in Figure 2 and indicated by Fleiss’ kappa. The results of the tests of proportions may be misleading, as they fail to reveal patient-level differences. Both data sources were equally likely to report a given diagnosis across the entire sample, but not necessarily for the same
patients. In aggregate tasks, such as calculating the denominator count of a CQM, the overall result is likely to be similar regardless of whether chart review or self-report is used. There is a clear and meaningful difference between these two sources, however, for individual patients, whether in the context of quality measurement or — more importantly—patient care. Of the 44 patients with myocardial infarction present in either chart review or self-report, for example, for only 20 is it present in both data sources. Out of all 380 diagnosis cases in our sample, fewer than half are shared amongst all three data sources.

In cases where a patient self-reported a diagnosis that was not present in chart review, it is generally difficult to discern whether the self-report was a false positive or the chart review resulted in a false negative. In other words, without further study it can’t be determined if patients were more likely to endorse diagnoses they did not have, or if they were reporting information that was missing from the OHSU medical records. In the opposite scenario, in which a patient did not report a diagnosis that was present in the chart review, the truth was sometimes easier to discern based on the level of evidence found in the chart,14 though the reasons why a patient would not report a diagnosis that is reported with high certainty in the chart remain difficult to understand or quantify without further research. Possible causes to consider include patient factors like limited patient health literacy and survey validity issues around the differences between lay language and clinical language.

Patterns of agreement across diagnoses

Across all the diagnoses, the highest level of agreement was for diabetes mellitus, which also had no significant differences of proportion for the pairwise tests between sources (though Cochran’s Q did show a difference across all three sources). This finding is possibly because this is a diagnosis associated with a high level of confidence (certain tests can be considered a gold standard in diagnosis) and a high level of actionability, meaning that it’s likely to be documented by providers and likely to be understood by patients. In the cases we reviewed where the three data sources were not in agreement, possible factors resulting in disagreement included the diabetes being very well-controlled (diagnosis factor), management by a non-OHSU provider (healthcare system factor), and research errors around data collection or survey design.

Heart failure and hypertension showed substantial agreement across the three data sources, though with more variability than for diabetes. The accuracy of hypertension diagnoses in EHRs is known to vary,15 perhaps because this is a diagnosis that often does require a certain accumulation of evidence, or because this is a diagnosis that is commonly managed by primary care providers. It is somewhat surprising that agreement was not higher across the data sources for heart failure. Error analysis showed that some of this disagreement can likely be attributed to the fact that heart failure may be conflated with related diagnoses, either in the chart or in patient self-report (e.g., congenital heart defect without heart failure). It may also be that patients with heart failure, owing to demographic differences or medical complexity, differ from other patient populations. In fact, the importance of health literacy in heart failure treatment and self-care has been highlighted in other work.16

Stroke and myocardial infarction showed the lowest agreement across the data sources, with Fleiss’ kappa for both indicating “moderate” agreement. As discussed above, disagreement between chart review and the problem list for these two diagnoses is at least partially due to the fact that prior acute events are not always added to the problem list. The large proportion of cases reported only in the patient survey—38% of potential myocardial infarction cases and 42% of potential stroke cases—suggests either that there is a substantial amount of relevant medical history that does not make it into the OHSU record, that patients are fundamentally misunderstanding their own medical histories, or some combination of the two.

Considering a role for patient self-report

The relationship between problem list data and diagnoses identified through manual chart review, i.e., that the former is largely a subset of the latter, is unsurprising. There is certainly value in developing approaches for encoding diagnoses present in unstructured data to make them more readily actionable. Patient self-report, however, provides information that is often entirely absent from the EHR, in either structured or unstructured data (22% of all diagnoses mentioned in any of the three data sources were found only in patient self-report). By definition, these diagnoses must be either true positives missed by the two EHR-based sources, or false positives, erroneously endorsed by patients or resulting from research error. More work is needed to quantify and characterize which cases are which, but it must be assumed that at least some of them are true positives, and that there is value in incorporating this information into the patient record. Our error analysis revealed that many of the diagnoses found only in the survey were from patients who received the majority of their care outside of OHSU, suggesting that in many cases we are observing fragmentation of medical information across the healthcare system. This is a recognized problem in healthcare,17 resulting in problems not just in quality improvement or research efforts, but at the point of care, when providers must make
clinical decisions without complete information. Transfers of care between providers and institutions are especially complicated from a medical record standpoint, and often the tasks of collecting and documenting complete medical histories from previous care settings is placed on already overburdened providers. Re-entering data that have already been encoded in structured data by other providers or at other institutions is an inefficient use of providers’ time. There are various approaches for overcoming this problem using technological and systems-based approaches, including health information exchange via regional health information organizations and standardized medical terminologies for medical data exchange.\(^{18-21}\)

In the era of the engaged patient, there may also be value in trying to incorporate patient self-report data into the EHR through properly validated and adjudicated means. There is already growing use of patient-reported outcomes in clinical care, quality assessment, and research; patients’ knowledge of their own medical status and history could potentially play a similar role, especially given that patients are already asked (often repeatedly) to provide information regarding their medical history. From a pragmatic standpoint, it is worth noting that although our initial survey development process was lengthy and a formal validation process is still required, once implemented this was an efficient way to collect data and, overall, was far less resource-intensive than the manual chart review process. Eliciting valid and reliable information from patients requires further study and must draw on existing work in psychometrics and patient-reported outcomes, but our findings strongly suggest that there is potential value in patient self-report if we wish to create complete, accurate, and current medical records.

Limitations

There are limitations to the generalizability of this work due to the focus on cardiology, the population definition (patients with cardiology visits in the past year), differences in documentation practices across institutions, and the limited type of clinical concepts included (only diagnoses), and possible response bias. We only invited patients with email addresses present in the EHR to complete the survey, which implies a certain degree of technological literacy and engagement with the healthcare system. Similarly, patients who completed the survey may have differed from those who did not in terms of engagement or literacy. Broader and more systematic work in this area is necessary to draw more generalizable conclusions about the differences between the three data sources. As mentioned in methods section above, we also recognize that there are inherent limitations and assumptions resulting from our decision to rely on problem list diagnoses instead of including encounter diagnoses. The sensitivity of the latter is generally higher, which could improve the sensitivity of secondary use cases relying upon structured data, though potentially at the cost of decreased precision.

Another limitation is that our survey, though informed by clinicians and medical literature, was not formally validated or assessed for reliability. There were also intentional tradeoffs in the survey design to increase accessibility of the survey items and decrease the overall time required to complete the survey, both of which may have led to decreased validity and granularity of information in the responses. As mentioned above, further work in this area should draw heavily on psychometrics approaches to instrument design and validation. The development of validated tools to assess patient medical history would be greatly beneficial.

Lastly, more formal error analysis is required to generate a more generalizable understanding of the factors that lead to differences between patient self-report, structured problem list data, and chart review data. A systematic approach looking at a larger representative sample of patients across multiple institutions would also tell us more about the degree of impact of the different causal factors, allowing prioritization of approaches to ameliorate these causes of disagreement. We intend to address these issues in future work.

Conclusion

Patient self-report and manual chart review both present an opportunity to identify significantly more clinical concepts than are present in structured EHR data. Moreover, the overlap between chart review data and patient self-report data is often limited and varies depending upon the concept of interest. Although the data quality problems associated with EHRs are increasingly well-accepted, manual chart review and abstraction is often considered to be the closest thing to a gold standard we have available, and the impact of data quality problems is not well understood. By comparing chart review data to patient self-report, we have demonstrated not only that chart review is sometimes erroneous, but also that the chart itself may be inaccurate, incomplete, or out-of-date. Further work is needed to determine when and to what extent patient self-report are reliable, as well as how best to collect such data in a valid fashion, but our initial results indicate that these data provide an opportunity for identifying data quality problems in the EHR as well as for potentially ameliorating those data quality problems.

Acknowledgements
Our thanks to Jayne Mitchell and James Mudd for their clinical expertise, to Michael Kahn, Stephen Persell, and Joan Ash for their guidance in methodology and synthesis, to Elizabeth Parker and Nelly Bostan for their patience and perseverance in data collection and chart review, and to the patients who participated in this study. This research was funded under National Library of Medicine K01 LM012738.

References

Motivation and Barriers to Using the Veterans Health Information Exchange: A Survey of Veterans Affairs ‘Superusers’

Kristen Wing, MA¹, Omar Bouhaddou, PhD², Nelson Hsing, PhD², Carolyn Turvey, PhD¹, Dawn Klein, LMSW¹, Joseph Nelson, MSA², Margaret Donahue, MD²

¹Department of Veterans Affairs, Office of Rural Health, Veterans Rural Health Resource Center, Iowa, City, IA, ²Department of Veterans Affairs, Office of Health Informatics, Veterans Health Information Exchange Program, Washington, DC

Introduction

The Department of Veterans Affairs (VA) Veterans Health Information Exchange (VHIE), formerly known as the Virtual Lifetime Electronic Record (VLER) Health Program, is a Veteran-focused portfolio of programs that gives VA and authorized Community Care Partners secure access to certain parts of a Veteran’s electronic health record for the purpose of care coordination across the health care continuum. VHIE established a technological platform to securely exchange health information between VA and the community, both at an organizational level (VA Exchange) and at the personal provider level (VA Direct Messaging). VHIE connects the Department of Veteran Affairs with care organizations through national networks (e.g., eHealth Exchange). Today, VA is connected to more than 314 eHealth Exchange members and will expand community provider access to Veterans health information by joining the CommonWell Health Alliance and Carequality national networks.

Each day, thousands of documents are sent and received by VA and community providers to coordinate the care of patients. However, VA provider adoption and utilization of the VHIE, specifically the Joint Legacy Viewer (JLV), remains low and far from the critical mass necessary to achieve desired impacts on Veterans care and on health care costs. As of March 2019, JLV usage reports showed only 17.3% of 311,920 VA staff with access to JLV were Active Users logging in to the system.

Provider adoption is one of the four pillars of the VHIE Program¹ and increasing VA provider utilization is key to program success for several reasons:

1. Veterans Health Administration (VHA) Directive 6371, released April 30, 2019, established that VA medical facilities “must implement the Veterans Health Information Exchange (VHIE) Program…which provides the capability of electronic health information exchange (eHIE)…to advance the delivery of Veteran-centered clinical care.” The directive requires implementing VHIE (JLV) technology and educating and training VHA staff on the use of the VA Exchange.

2. The VA MISSION Act of 2018 specifically addresses health data exchange between VA and the private sector and opened the door for a shift to opt-out consent (no signed consent required) for Veteran participation in VHIE. This policy change will drastically increase the amount of Veterans health information available for electronic exchange and will create internal and external expectations that (all) VA providers use the VA Exchange to manage Veterans care.

3. Most enrolled Veterans (80%) have other health insurance (e.g., Medicare, Medicaid, Tricare, private insurance) and are likely to receive care outside of VA. This rate could be higher for rural Veterans as community care may be more accessible than VA care.

4. “Repetition of medical services by providers is one of the major sources of healthcare costs,” and use of HIEs to coordinate patient care can reduce redundant and unnecessary medical tests and lead to significant cost savings⁶⁻⁸.

In 2016, the VA Office of Rural Health (ORH) and VHIE program office began collaborating on a social marketing (communications) project to increase VA provider adoption and utilization of the VHIE. The VHIE Superuser Survey, conducted in 2018, was designed to help gain a deeper understanding of VA employees demonstrating the “model” behavior – accessing and using the VA Exchange at a high rate. The survey investigated VHIE utilization patterns and user perceptions about the VHIE and its use, as well as data of interest to the VHIE program regarding the user experience, including system access, use case(s), data quality, training, and barriers and facilitators to using the VHIE.
The survey also sought to discover psychographic information about “superusers” to better understand them as a consumer of HIE technology.

Survey findings will inform the development of a strategic communications campaign to increase VHIE adoption and utilization by VA providers and staff. In addition, knowledge gained from the VHIE Superusers survey will be used to help prioritize enhancements to the program, reduce and remove barriers to VHIE adoption and utilization, and improve the user experience.

HIE adoption is also an issue facing community health care organizations and barriers are similar to those experienced by VA(9-13). This work serves to add to existing knowledge about the motivators, facilitators, and barriers to HIE adoption; and the use of social marketing in interventions to influence provider behavior change.

Methods

Social marketing uses commercial marketing concepts and techniques to promote voluntary behavior change, and has been employed by international and national public health interventions for decades(14, 15). The social marketing process includes nine process elements completed across three phases: I) Research and Planning, II) Strategy Design, and III) Implementation and Evaluation(16). It is “an invaluable referent from which to design, implement, evaluate, and manage large-scale, broad-based behavior-change focused programs(17).” Though patients (consumers) are generally the intended target, social marketing may also be used to influence behavior change in health care providers(18-20).

The VHIE Superuser Survey was developed as the research vehicle to collect the quantitative and qualitative data necessary to complete the consumer analysis, market analysis, and channel analysis elements of Phase I. Survey questions were designed to elicit information regarding the VHIE user experience as well as to answer six strategic questions posed by the consumer-based health communications (CHC) process. Answers to these questions (below) provide insight into the consumer’s reality and “lead to communications that are relevant, meaningful, and compelling to the audience(21).”

1. Who will the target consumer be and what are they like?
2. What action should the target take as a result of communication?
3. What reward should the message promise the consumer?
4. How can the promise be made credible?
5. What communications openings and vehicles should be used?
6. What image should distinguish the action?(21)

To confirm the selection of the VHIE Superuser as an appropriate model of the desired behavior, questions based on Theory of Planned Behavior constructs were included to assess target audience members’ intention to perform the desired behavior (VHIE utilization in the workflow), their attitude toward the desired behavior, subjective norm, and perceived behavioral control.

VA Research Electronic Data Capture (REDCap), a web-based application for building and managing online surveys, was used to create open-ended, single-select, multi-select, and Likert Scale questions. VHIE and ORH staff, and VHIE Rural Health Community Coordinators tested the survey and provided feedback regarding question clarity and survey completion time. The final VHIE Superuser survey consisted of 32 questions and addressed four domains: Access, Utilization, Data Quality, and User Experience.

VHIE analytics were used to identify VA medical centers/healthcare systems retrieving documents from external sources at high (>14,000 patient record views/facility) and low (<2,000 patient record views/facility) rates. Nine sites were selected for participation in the VHIE Superuser survey; four designated as rural and five as urban.(22) [Comparison of rural and urban is of interest to the ORH as rural Veterans may be more likely to obtain care in the community due to distance/travel time to a VA facility, and ORH supports VHIE Community Coordinators at 56 rural VA Medical Centers.]

As this research was focused on the VHIE “consumer” and identifying current behavior and a model of the desired behavior, a purposive sample was established consisting of 162 VHIE users with the highest number of patient record
views at their site in the previous six months. Email addresses for the target Superusers were confirmed in the VA Outlook Global Address List; five were “undeliverable” reducing the total survey population to 157.

The initial survey email was sent July 17, 2018 and survey reminder emails were sent July 24th, July 31st, and August 7th. To increase the survey completion rate (only 15.3% as of 7/31/18), content and formatting changes were made to the final two email reminders. Emails were personalized with information specific to the recipient, such as their position title and home facility; and persuasive language and visually-appealing graphic elements were added.

Results

The final survey response rate of 52.2% was calculated using a sample size of 82. The sample was comprised of 36 participants from four rural sites (44%) and 46 participants from five urban sites (56%). Among all survey respondents there was wide variation in years worked at VA, with the shortest duration six months and the longest tenure slightly more than 38 years. The mean time employed with VA was 10.6 years. Nurses responded to the survey invitation at the highest rate (39.0%), followed by physicians (26.8%) and pharmacists (13.4%). Respondents who selected the “Other” category (23) included Psychologists (6), Social Workers (4), and Vocational Rehabilitation Counselors/Specialists (4).

It is important to note survey participants responded to questions based on their experience with either one or both VHIE user interfaces: VistAWeb or the Joint Legacy Viewer (JLV). VistAWeb, an intranet web application providing read-only access to individual patient electronic health records (EHR) was implemented in 2010 and served as VA’s first health information exchange (HIE). To modernize the system and meet Federal mandates for interoperability, JLV was deployed to all VA Medical Centers in 2014. JLV, co-developed by VA and the Department of Defense (DoD), connects VA to the private sector through national networks and enables users to view comprehensive electronic health records from all VA, DoD, and community health information exchange partner facilities where a Veteran receives care. An important difference between JLV and VistAWeb is that VistAWeb allows for aggregated data views that combine internal and external data in single data displays (e.g., allergies together, medications together, problems, etc.). VistAWeb, which was scheduled to be decommissioned in September 2017, was still available and in use during the survey period.

Survey Domain #1: VHIE Access

To better understand the “how” and “how often” aspects of VHIE access, and to identify any access issues, questions in this domain concerned the interface used to view Veterans non-VA (community) health information, frequency of access, and perceptions of ease in accessing the VHIE. Respondents were also asked to share open comments about VHIE navigation in general.

Survey participants were asked to select all methods used to access Veterans non-VA health information. More than half of respondents (52.6%) reported using JLV, and slightly more than a third (35.3%) used VistAWeb; 61.4% selected both JLV and VistAWeb. This may indicate that VHIE Superusers were early adopters of VistAWeb and that many have been successfully transitioned to JLV.

Almost three-quarters of Superusers (73.4%) access the VHIE daily, and overall ease of accessing data in the VA Exchange was rated “Neutral” (3.03) on a 5-point Likert Scale. Regarding access, 82% of survey respondents (n=67) provided comments about VHIE navigation and the user interface. Qualitative analysis of text responses revealed three predominate themes: connectivity, functionality, and a preference for VistAWeb over JLV. For both connectivity and functionality, the speed (slowness) of the system was an overarching theme. Almost one quarter of respondents (22.4%) used the word “slow” in regard to connecting and logging into the VA Exchange as well as loading documents (patient records), and “time-consuming” was referenced five times. Inaccessibility to the VHIE because the “system was down” was cited by 11 respondents, and 10 specifically referred to issues with the log in process (two-factor authentication). Almost one-fifth of survey respondents (19.4%) mentioned the transition from VistAWeb to JLV, compared the functionality and user friendliness of the two interfaces, or declared a preference for the system they had already been trained to use and are accustomed to using in their workflow.
**VHIE Access Themes – Sample of Supporting Superuser Quotes**

**Theme #1: Connectivity issues prevent access to the VHIE.**

“Becomes problematic when sites are not available or when the JLV goes down as there is no other alternative to get the info I need.”

“JLV is not quickly accessible to log in and sometimes goes down.”

**Theme #2: Functionality and navigation inhibit utilization of the VHIE (JLV).**

“Too limited – can’t search, often have to open multiple single entries to try to find what you’re looking for. Often need information like a test result that is not retrievable (not the most recent or several years ago).”

“I find JLV somewhat difficult to navigate to [location] information I am seeking.”

**Theme #3: Superusers were comfortable using VistAWeb and prefer it over JLV.**

“When Vista Web was available, the information I was able to access was more useful and decreased the time to find pertinent data that would help the clinical team make treatment decisions…”

“I prefer to use VistAWeb over JLV - some of that may be that I have more experience with VW, but it also allows us to access information not available in JLV.”

**Survey Domain #2: VHIE Utilization**

The second survey domain assessed utilization of the VHIE as influenced by Veteran patient characteristics, community health data types searched for in the VA Exchange, and how community data was used in the work setting. Participants were asked to check all responses that applied to the statement, “I access VLER/VHIE when working with a Veteran who...” Selection of [Veteran] “has self-reported care in the community” prompted a follow-up question regarding the type of community care reported by the Veteran.

Sixty-eight percent of respondents reported using VHIE to look up patient records when working with a new enrollee or patient transferred from the DoD, and 61.8% accessed the VHIE when a Veteran self-reported community care. Interestingly, 36.8% reported a Veteran-initiated request to look up their health record. Almost half of the respondents (48.7%) selected “Other Patient Type” and provided varied text responses. Most commented on clinically-specific situations, but use of the VHIE when a Veteran referenced medication(s) prescribed by a provider outside VA emerged as a theme.

<table>
<thead>
<tr>
<th>Community Imaging</th>
<th>52.6%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community ER Visit</td>
<td>47.4%</td>
</tr>
<tr>
<td>Community Specialty Care Tx</td>
<td>47.4%</td>
</tr>
<tr>
<td>Community Labs</td>
<td>43.4%</td>
</tr>
<tr>
<td>Community Chronic Care Tx</td>
<td>40.8%</td>
</tr>
<tr>
<td>Veterans Choice Program</td>
<td>31.6%</td>
</tr>
</tbody>
</table>

**Figure 1.** Types of community care reported by Veterans to VHIE Superusers (n=76).
As shown in Figure 1, more than half of respondents (52.6%) reported Veterans referencing receiving diagnostic imaging in the community. Specialty care treatment and visiting community Emergency Rooms (ERs) were both reported by 47.4% of respondents, followed by labs in the community (43.4%) and community care for a chronic condition (40.8%). Less than one-third mentioned Veterans acknowledging health care from a community provider through the Veterans Choice Program.

In terms of interest in specific data categories, nine of the 13 listed data types were reported by more than half of survey respondents (=>51.3%) (Figure 2). More than three-quarters of respondents searched for problems/medical history (78.9%) and lab results (77.6%). While viewing all data types has the potential to impact and improve care coordination, more than two-thirds of respondents reported searching for lab results (#2), imaging results (#3), medications (#4), and procedures (#5); which may contribute to cost reduction through eliminating duplication.

**Figure 2.** Data types searched for in VHIE (n=76).

**Figure 3.** VHIE Superuser data use cases (n=76).
Responses to the question regarding utilization of Veteran health information obtained through the VA Exchange are shown in Figure 3. The top five uses, (1) reviewing treatment course for chronic illness, (2) reviewing community hospitalizations/ED notes, (3) preparing for scheduled appointments, (4) coordinating care, and (5) managing medications were reported by more than half of respondents.

Survey Domain #3: VHIE Data Quality

The third survey domain assessed Superusers perceptions of the quality of the health information (i.e., data accuracy and completeness) received from the VA Exchange. The overall mean score for data quality was 3.85 on a 5-point Likert Scale where “1 = Very Poor” and “5 = Very Good.” Close to two-thirds of respondents (63.9%) rated VHIE data quality as either “Very Good” or “Somewhat Good.”

The top three responses selected for the question regarding data quality issues experienced when using VHIE were (1) incomplete data (59.3%), (2) duplicate data in multiple locations (37.3%), and (3) “other” data quality issue” (33.9%). Five survey participants who selected “Other data quality issue” reported that there were “no problems” with data quality, and one respondent added, “other than not every facility can/has signed up to play with us.”

When asked about specific data not currently available that would be helpful in their work, several respondents referred to progress notes - either more or better (from the community). Other responses included a list of patients’ current providers, surgical history, and psychological evaluations. Two items suggested by multiple respondents were current and past appointments (VA and community) and prescriptions written by community providers. It should be noted that this was not a required question, yet 10 respondents voluntarily responded with “not applicable” or “none,” indicating they were able to access the information to perform their work.

Survey Domain #4: VHIE User Experience

The final domain of the VHIE Superuser survey collected information regarding the VHIE user experience and perceptions of benefits and drawbacks. In addition, motivation for VHIE use, facilitators and barriers affecting utilization, and VHIE training received were addressed in this domain.

Sixty-six survey participants shared their perception of the benefit(s) of using the VHIE in their work. Qualitative analysis of the text responses revealed three predominate themes: coordination and continuity of care, access to outside information, and better patient care. Expediting health record sharing (i.e., the process to obtain Veterans health information from other VA and non-VA providers) and eliminating duplication (e.g., tests, treatment) were also strong themes.

VHIE Benefits Themes – Sample of Supporting Superuser Quotes

Theme #1: VHIE use helps provide coordination and continuity of care.

“Accessing digital information in real time is invaluable in providing coordinated, comprehensive care.”

“Helpful tool to increase continuity, efficiency, and thoroughness for patient care.”

Theme #2: VHIE use provides access to helpful health information from other VA and non-VA providers.

“Being able to gather a complete medical picture of the Veteran from both VA and non-VA care to provide the best possible care possible.”

“It allows me to follow the Veterans who receive their primary care outside the VA - to follow labs, follow medication dosing, and see the education that has been provided by non-VA providers.”

Theme #3: VHIE use leads to better patient care.

“Allows me to better evaluate the Veteran’s needs, previous/current care, medical stability...”

“Outside records greatly benefit Veteran care.”

Sixty-five survey participants shared text comments about their perception of the drawbacks of using the VHIE. More than half of respondents (55.4%) cited technical issues as a drawback. Within this theme, slowness of the system was
the greatest drawback (30.8%), followed by inaccessibility due to the system (JLV) being down (9.25%), log in/authentication issues (7.7%) and navigation (7.7%). The second major theme concerned the lack of data available from community partners (18.5%). The third theme, “no drawbacks to VHIE use,” was mentioned by 13.8% of respondents to this question.

**VHIE Drawbacks Themes – Sample of Supporting Superuser Quotes**

**Theme #1: Technical issues are a drawback of VHIE utilization.**

“The process is so slow at times that it is not always possible to utilize in the context of a standard clinic appointment time. Just the process of logging into JLV can be cumbersome…”

“System is frequently down or access is slow, also amount and type of information is limited.”

**Theme #2: Lack of information available from community providers is a drawback of VHIE utilization.**

“Unfortunately, not all community providers are accessible in VHIE.”

“Limited facilities who participate and notes at outside facilities are often sparse.”

**Theme #3: There are no drawbacks to using VHIE in my work.**

“There is no drawback that I can think of.”

“None.”

When Superusers were asked what motivated them to use the VHIE, the overwhelming choice selected by survey participants (84.5%) was “desire to improve care coordination” A formal directive from leadership motivated less than a quarter of respondents. Fifteen respondents chose to report an “other” motivation, which included participation in a JLV test, efficiency, and for claims processing.

User Experience domain questions also intended to discover the perceptions of facilitators and barriers to utilizing the VA Exchange. Survey participants were asked to select all applicable items. Of sixty-seven respondents, more than three-quarters (77.6%) reported that ease of accessing the VHIE in the work setting facilitated its use. Other top facilitators included facility support for VHIE use, and the ability of support staff to access the VHIE.

Through their responses to questions in the VHIE Access domain, Superusers established clear themes regarding barriers and echoed them in the User Experience domain: connectivity, functionality, and a preference for VistAWeb over JLV. Ninety percent of respondents reported “slow connection or system not available” as a barrier, almost two-thirds (63.4%) perceived the “time consuming” aspect of VHIE use as a barrier, and one-third felt VHIE use interfered with the workflow. Text responses for “Other” barriers were similar to already established technology-related themes concerning the log in process, navigation, and functionality.

Eighty-three percent of survey participants indicated they had been “self-trained.” The VHA Office of Health Informatics (OHI) launched a comprehensive communications campaign to alert all VA Computerized Patient Record System (CPRS) users about the transition from VistAWeb to JLV. Although technological limitations prevent definitive attribution, the number above would suggest many users were self-trained using the JLV videos promoted in OHI communications. Other training methods, such as group training or one-on-one training were reported at significantly lower rates: 16.9% and 2.8% respectively. Those respondents who selected “Other VLER/VHIE training” (9.9%) referenced the JLV test email, word-of-mouth, and learning by trial-and-error.

It should be noted rural Superusers (69.4%) reported far more facilitators of VHIE use than their urban counterparts (30.6%). Facility support for VHIE use (83.3%), ability of support staff to access VHIE (82.4%), and effective training (75%) were the top three facilitators. VHIE Community Coordinators at rural sites were considered a facilitator by 71.4% by respondents. Rural Superusers (41.7%) also reported less barriers to VHIE utilization than urban (58.3%).

To further understand why Superusers are utilizing the VHIE at such a high rate, survey participants were asked to respond to a series of statements assessing Theory of Planned Behavior constructs. This theory, often employed by patient-level health behavior change interventions, is used to predict the likelihood of an individual performing a behavior based on attitude, subjective norm, and perceived behavioral control. On a 5-point Likert Scale where “1 =
Strongly Disagree” and “5 = Strongly Agree,” respondents expressed a positive attitude about the VA Exchange and its use, support from their peers for using VHIE, and confidence to successfully use the VHIE. Each statement scored 4 or higher except for the statement regarding peer use of the VHIE (3.47). These results indicate the Superusers are a good “model” of the desired behavior (VHIE adoption and utilization).

![Figure 4](image-url)

**Figure 4.** Theory of Planned Behavior Constructs and VHIE Superuser Behavioral Intent (n=72).

**Discussion**

Whether working in a rural or urban facility, those who currently use the VA Exchange most, the Superusers, share commonalities. They have a positive attitude toward VHIE, despite technical barriers impeding or preventing use. Their shared desire to deliver high quality, coordinated care to Veterans and their ability to access the VHIE throughout the work setting motivate and facilitate VHIE utilization. The Superusers believe there are benefits of incorporating use of the VA Exchange in their workflow and have identified use cases for better and more efficient care delivery. They feel that while data quality is good, gaps still exist and HIE participation by more community providers would help provide salient missing information. As 70% of the survey respondents have been with VA five years or more, they are experienced with and more comfortable using VistAWeb, but through self-directed training can develop the necessary knowledge and skills to successfully transition to JLV.

As key consumers of Veterans health information, Superusers have expressed a clear intent to look outside VA for information to coordinate Veterans care. They routinely view community medical histories, lab and imaging results, medications, and procedures and notes. They have found the value of VHIE use to co-manage care of Veterans with chronic conditions and community hospitalization or emergency department admissions. They gather information to prepare for appointments and often review discharge summaries, immunizations, and allergies.

The VHIE Superuser Survey effectively captured data and psychographic information to create a vivid picture of the model VHIE user and their reality. Survey findings and answers to the six CHC strategic questions and are currently being used to develop a communications campaign promoting the benefits of VHIE use, facility/peer support for VHIE use, and the availability of self-directed training videos to quickly gain the skills necessary to begin accessing and using the VHIE in the clinical setting (social marketing Phase II). The campaign is scheduled to launch at 56 rural VA Medical Centers in late summer 2019. Throughout the campaign JLV training video completion will be monitored; and Active JLV Users, Patient Record Views, and Unique Patients Viewed will be measured. Program evaluation (Phase III) will measure the communications intervention ability to influence behavior change and what, if any, role the self-directed JLV training played in technology adoption and utilization.
Conclusions

While the MISSION ACT of 2018 will expand access to community care for Veterans, and an opt-out policy change will dramatically increase the amount of Veterans health information to be exchanged between VA and community providers, meeting the full intent of the law and VHA Directive 6371, and maintaining Veterans trust in VA, requires a greater number of VA providers actively using the VHIE. As the opt-out consent policy for VHIE goes into effect in October 2019, Veterans will be asking their VA providers if they are viewing electronic community health records as part of standard care.

Results of the VHIE Superusers Survey are consistent with previous study findings indicating those who use the VA Exchange are motivated to do so by a desire to provide high quality, coordinated care for patients, and that technical barriers may prevent users from incorporating VHIE into their workflow and encouraging others to do the same (23). Video training, available for viewing at the user’s convenience, provides an attractive and effective alternative to group training or one-on-one training; especially for individuals who perceive time for skills-building as limited.

Employing the social marketing process to address low HIE utilization may be an effective approach to influencing and changing provider behavior, but requires planning and research to collect the quantitative and qualitative data necessary to understand the target audience and their reality. As the VA transition to a new EHR looms on the horizon (as it does for many community providers), an effective evidence-based communications program facilitating new health IT adoption has application inside and outside VA.

Acknowledgements

The authors would like to thank the VHIE Superusers, VHIE Community Coordinators, and VHIE program office team for their time and assistance with this study. This work was funded by the Department of Veterans Affairs Office of Rural Health and conducted in collaboration with the VA Office of Health Informatics, VHIE Program Office. It received Quality Improvement determination from the VA Maine Research and Development Committee and met compliance with VA National Center for Organizational Development, Organizational Assessment Sub-Committee and VA Office of Research and Development guidelines for surveying VA employees for VA operational and research purposes.

References


922
User engagement with web-based genomics education videos and implications for designing scalable patient education materials

Julia Wynn, MS, CGC1,a, Wei Wei, PhD2,a, Xinhang Li, MS2, Yat S So3, Suzanne Bakken2,4, PhD, RN, Chunhua Weng, PhD2,b, Wendy K. Chung, MD, PhD1,5,b

1Department of Pediatrics, Columbia University Irving Medical Center, New York, NY
2Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA
3Irving Institute of Clinical and Translational Research, Columbia University Irving Medical Center, New York, NY, USA
4School of Nursing, Columbia University Irving Medical Center, New York, NY, USA
5Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA

a Ms. Wynn and Dr. Wei contributed equally to this work
b Dr. Chung and Dr. Weng contributed equally to this work

Abstract
Genomic medicine has created an urgent need for scalable genomic education. One promising approach is self-guided learning platforms. Understanding how these platforms are used is critical to guide their effective development and implementation. This study contributes a log-based method to study user engagement with online genomic educational videos among participants in a genomic screening study. We collected baseline demographics, logged participant usage and compared pre- and post-education genomic knowledge. Participants (N=390) who chose website access differed from those who declined access (N=81) and were more likely to be non-Latino, English speaking, younger, and have higher educational attainment. Only 45% who accessed the website viewed at least one video. The average video exposure time was 12 minutes. Longer exposure was not associated with an improvement in the user’s genomic knowledge. Our study and future studies of user analytics should be used to guide the development of effective, scalable genomic education methods.

Introduction
The integration of genomic medicine into research and healthcare has created a need for effective, efficient, and scalable genomic education. Traditionally, genomic education is provided by a healthcare provider, often a genetic counselor, in a one-on-one, face-to-face session. With the growing practice of genomic medicine, this model is not sustainable due to workforce limitations, including lack of access to genetic counselors1, limited expertise among nongenetic providers2,4, and stringent time constraints for all healthcare providers. Additionally, in the research setting, there is growing consensus that researchers have an obligation to return individual genetic research results though they may not have the knowledge or infrastructure to provide the necessary genomic education5. These changes and accompanying challenges raise the need to not only develop alternative genomic education methods, including web-based video methods, but also approaches to measuring the use and effectiveness of these platforms.

Studies have begun to demonstrate the potential effectiveness of alternative genomic education methods. Several studies have shown videos and interactive web-based methods to augment or provide equivalent genomic education to traditional genetic counseling. Web-based tools are also positively endorsed by providers and patients6-12. A study of non-patient users demonstrated improved genetic knowledge after video education on genomic sequencing13. Studies have also demonstrated videos to be effective for non-genetic medical education and associated with improved knowledge, retention of information and patient satisfaction14-16. With the exception of one study that measured the amount of time a participant spent watching the video15, in all of these studies, participants’ usage of the tool was known because they accessed the tool in the clinical setting or because they followed prompts to the end of the web-based education; no more detailed user analytics were collected. There is one study using public YouTube videos on hereditary cancer syndromes that collected frequency of views but had only limited information on a fraction of the people who viewed the video and had no measure of effectiveness or satisfaction17.


While emerging research demonstrates the potential utility of web-based genomic video education, these studies have lacked data on how individual participants interact with these platforms and content, and how this use is or is not associated with a measure of effectiveness (e.g. change in genomic knowledge, participant satisfaction, etc.). A more complete understanding of participants’ use including the minimal user engagement is needed, especially when the education materials are self-guided and administered outside of a clinical setting.

We developed a log-based analysis to measure individual engagement with web-based education videos among the participants enrolled in a study of genomic screening for Ashkenazi Jewish and Latino individuals. We also explored the associations of participant usage of the videos with demographics, change in genomic knowledge and participant perceived benefits of the educational videos. Our study informs how user data can be collected and should be used to guide the development, assessment and use of future genomic education methods in the research and clinical settings.

Methods

We conducted a secondary analysis of IMAGene participants’ study website and video usage. The IMAGene (Individualized Medicine through the Application of Genomics) (www.imagenestudy.org) study is part of the Electronic MEdical Record and Genomics (eMERGE) consortium. eMERGE is currently in Phase III, and in this phase approximately 25,000 people enrolled in parallel consortium studies are receiving results from genomic sequencing. The IMAGene study is an arm of the Columbia eMERGE study which offered genomic screening using targeted gene sequencing/genotyping to adults who self-identified as Ashkenazi Jewish or Latino/a, who could read and speak English or Spanish and had received care at NewYork Presbyterian – Columbia University Medical Center. The restricted IMAGene study sample addressed the larger eMERGE 3 goal of assessing the ability to interpret genomic data from minority populations. The IMAGene cohort is a non-randomized, convenience sampling of individuals who met study eligibility criteria, actively invited through letters and phone calls or passively invited through flyers and community engagement events. Participants were invited to complete the study via a website, on paper through the mail, by phone or in-person. This analysis focuses on those participants who chose to complete the study and genetic education through the website. The study was approved by the Columbia University Medical Center Institutional Review Board.

Following informed consent, all participants completed a baseline questionnaire, which included demographic questions and a 15-question genomic knowledge scale (GKS). The GKS included questions that were adapted from published scales. After completion of the baseline questionnaire, participants who elected genetic education via the website were invited to view the content of the study website including the study videos. Participants accessed the secure study website at their convenience using their email address and an assigned password. The participant self-determined when they had sufficient genetic education and were not required to view or watch any content on the website. They were then prompted to complete the GKS again (post-education GKS). Participants had two attempts to answer 12 or more of 15 questions on the post-education GKS correctly. If they were unable to do this, a genetic counselor contacted them before they were able to proceed with the study and have genomic screening.

We developed educational videos and documents for this study to provide self-guided genomic education through the study website https://www.imagenestudy.com. The videos and website content were developed by a team of geneticists, psychologists, genetic counselors, research assistants and community members over the course of 12 months. The videos were reviewed on three occasions by the development team and iterative improvements were made. The website hosts eight videos (Table 1); videos 2-7 were revised to improve clarity on January 27, 2017. Video 5 was removed after March, 2017 because of a change in the study protocol, and analysis of this video is not included in this study. In this analysis, we present user data from the old and new videos together unless otherwise specified. The video topics included genes, inheritance and genetic mutations, genomic screening for genetic conditions, interpretation of different types of screening results (e.g., personal genomic results and carrier screening) and personal, familial and reproductive implications of screening results. The average video duration for the English language videos was 3.5 minutes. The total time to view all videos was 26 minutes. The videos’ Felsch-Kinkaid reading grade level ranged from 9.0-12.0 and 6.4-10.6 when the genetic terms that were defined in the video were removed. The Spanish language video covered all topics and was almost 10 minutes long. It was viewed by only four participants, and therefore no analysis for this video was completed.
Participants who elected not to have website access received a DVD of the website videos and paper materials that included the same information as that on the website. Since we were unable to measure usage of these offline materials, these participants were not included in our usage analysis. All participants were notified of the option to contact study personnel or the study genetic counselor by phone or email.

The website was accessed by participants on their own devices at their convenience from July, 2016 to August 2018 and unobserved by study personnel. We logged and measured user specific data including website access and duration and type of video viewed and performed a retrospective analysis of the log files of the website with focus on log data for video playing patterns. The atomic unit of the log files is event. Each event corresponds to a type of action. The analysis was performed using Python2.7, and the scripts are available upon request.

We measured the number of participants (i.e., users) who accessed the website, the total number of times they accessed the website, and the length of these accesses. Website exposure is the sum of all the website accesses for each participant. We measured the number of videos users accessed, number of videos users completely viewed, and the total time of video access. Video access is defined by the observation of a “video playing” event. Repetitive accesses to one video counted as one access. The video exposure time is the total length of time for all video accesses including repetitive views. The videos are composed of pages of content which are similar to animated slides in Microsoft PowerPoint, and this internal structure is referred to as a “page” in this analysis. The last page of each video displays study contact information without audio. Video completion was defined as reaching the last page of video. The events recorded in the log files were segmented into sessions to analyze users’ actions on each video. Each event is associated with a page view ID. We created sessions by grouping events based on page view ID prefix, and a session is dedicated to the access of one video from one user.

To enable the evaluation of associations between exposure time and participant characteristics, participants’ video exposure times were divided into three categories: (1) minimal - those who had web access but had less than 120 seconds of video exposure time, (2) moderate - those who had 120 to 600 seconds of video exposure time and (3) maximal - those who had over 600 seconds of video exposure time.

Summary statistics are presented in frequencies, means and ranges. Unadjusted analysis included Chi squared analysis and Fisher exact tests to evaluate categorical variables, and two sample t-tests and ANOVA to evaluate continuous variables. Forward multiple regression models were built to further examine the relationships between video exposure time and change in GKS score. In the simple analysis, demographic and baseline variables associated at an alpha <0.1 with predictor, video exposure time, and the outcome, change in GKS score, were defined as confounders. The final models adjusted for the identified confounders. We completed the analysis for old and new video exposure times separately and combined exposure times. The results did not differ, so we present the combined analysis. Statistical analysis was complete in SAS 9.4. Given the exploratory nature of this analysis, we did not assign a threshold of statistical significance.

Table 1. Educational videos on the IMAGene Website.

<table>
<thead>
<tr>
<th>Video ID</th>
<th>Video Title</th>
<th>Video Length (minutes)</th>
<th>Language</th>
<th>Flesch-Kinkaid Reading Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Old</td>
<td>New</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Introduction to Genetics</td>
<td>3.50</td>
<td>3.50</td>
<td>English</td>
</tr>
<tr>
<td>2</td>
<td>Introduction to Genomic Screening</td>
<td>3.53</td>
<td>2.77</td>
<td>English</td>
</tr>
<tr>
<td>3</td>
<td>Personal Genomic Risk</td>
<td>2.03</td>
<td>3.68</td>
<td>English</td>
</tr>
<tr>
<td>4</td>
<td>Carrier Screening</td>
<td>2.75</td>
<td>3.53</td>
<td>English</td>
</tr>
<tr>
<td>5</td>
<td>Pharmacogenetics*</td>
<td>5.38</td>
<td>2.43</td>
<td>English</td>
</tr>
<tr>
<td>6</td>
<td>Possible Genomic Screening Results</td>
<td>2.95</td>
<td>7.07</td>
<td>English</td>
</tr>
<tr>
<td>7</td>
<td>Reproductive Options</td>
<td>1.53</td>
<td>3.58</td>
<td>English</td>
</tr>
<tr>
<td>8</td>
<td>Introducción a la genética y IMAGene</td>
<td>9.82</td>
<td>9.82</td>
<td>Spanish</td>
</tr>
</tbody>
</table>

Videos were updated on January 27, 2017. “Old” indicates the videos before the update, and “New” indicates the updated videos.

* Video 5 “Pharmacogenetics” was removed from the website in March 2017.
Results

A total of 471 participants were enrolled, and 390 elected access to the website. The participants who elected access differed demographically from those who chose not to access the website. Participants who elected website access were more likely to be non-Latino/a, US born, English speaking, younger than 45 with a greater-than-high-school education, and privately insured (Table 2). Only 14 participants without website access proceeded with the study, and therefore it is difficult to assess differences between these two groups for these variables. The participants who had access to the website performed better on the 15 pre-education GKS (no website access: average 9.6 range 0-15, website access: average 11.8, range 2-15; p-value <0.0001). There was no difference in the change in the GKS from baseline to post education between the two groups (no website access: average 5, range 1-5, website access average 1 range -6-11; p-value 0.19).

Table 2. Demographics of participants who requested website access and those who declined. N (n is noted next to the variable if data are missing), column percentage. Chi square and Fisher

<table>
<thead>
<tr>
<th>Baseline Variables</th>
<th>No Website Access</th>
<th>Website Access</th>
<th>p-value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>81</td>
<td>390</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>8</td>
<td>104</td>
<td>0.0012</td>
</tr>
<tr>
<td>&lt; 45 (n=79)</td>
<td>11</td>
<td>192</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Latino&lt;sup&gt;b&lt;/sup&gt; (n=79, n=383)</td>
<td>63</td>
<td>221</td>
<td>0.0002</td>
</tr>
<tr>
<td>Private Insurance (n=60, n=317)</td>
<td>13</td>
<td>197</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>≤ High School (n=79, n=383)</td>
<td>55</td>
<td>137</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>English Speaking</td>
<td>26</td>
<td>323</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Born in US (n=79, n=384)</td>
<td>23</td>
<td>250</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Web Q&lt;sup&gt;c&lt;/sup&gt;</td>
<td>4</td>
<td>291</td>
<td></td>
</tr>
<tr>
<td>Completed post-education Q</td>
<td>14</td>
<td>320</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Abbreviations: questionnaire (Q), genetic knowledge scale (GKS)

<sup>a</sup> comparison of participants with and without website access
<sup>b</sup> Non-Latino participants identified as Ashkenazi Jewish
<sup>c</sup> 4 participants who completed the questionnaire electronically and completed study enrollment before the website was live.

Of the 390 participants who had access to the website, 303 (78%) accessed it a total of 1166 times. The average number of website sessions per user was 3.8 with a range of 1 to 32 times. The average total time of website sessions per user was 5.8 minutes with a range of 0-50.5 minutes (Figure 1).
Of the 303 users who access the website, 136 (45%) accessed a video in over 617 website sessions. The average video exposure time per user was 13.5 minutes with a range of 1 second to 34 minutes for old videos and 12.5 minutes with a range of 1 second to 45 minutes for new videos (Figure 2 a,b). The combined (old and new videos) average video exposure time per user was 12 minutes with a range of 1 second to 50 minutes. For the 136 participants who accessed a video (old or new), the average number of videos accessed was 3.8 with a range of 1-7 (Figure 3a,b). The viewing patterns of the individual new videos were assessed. Evaluation of the user viewing patterns showed that the greatest viewer drop-off occurred in the first five seconds. Users who spent more than five seconds on the video typically completed the video to the final page (data not shown).

**Figure 2ab.** Histogram of video access time for the old (a) and new videos (b).

**Figure 3ab.** Histogram of number of videos accessed for the old (a) and new videos (b).
The frequency of users by video exposure time were: (1) minimal (n=269), (2) moderate (n=53), and (3) maximal (n=68). There were a few differences in the demographics by exposure time; participants who were English speaking and completed the study questionnaires online were more likely to have moderate to maximum video exposure time. People who did not complete post-education questionnaire had lower video exposure time (Table 3). Video exposure times were associated with baseline 15 question GKS score (average 11.5 (range 2-15), 11.9 (8-15), 12.6 (6-15), respectively, p-value 0.003).

In the unadjusted analysis there was no difference in the change from baseline to post education in GKS across video exposure groups (Figure 4). After adjusting for the confounders of baseline GKS and insurance status, there remained no association between video exposure time and change in GKS score. In the adjusted analysis, the only predictor of change in GKS score was a lower baseline GKS score (Table 4).

### Table 3. Demographics of participants who had access to the website stratified by video exposure time. N (n is noted next to the variable if data is missing), column percentage. Chi squared and Fisher exact.

<table>
<thead>
<tr>
<th>Baseline Variables</th>
<th>&lt;120</th>
<th></th>
<th></th>
<th>120-600</th>
<th></th>
<th>&gt;600</th>
<th></th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>269</td>
<td>53</td>
<td>68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>73</td>
<td>10</td>
<td>21</td>
<td>31%</td>
<td></td>
<td></td>
<td></td>
<td>0.32</td>
</tr>
<tr>
<td>&lt; 45</td>
<td>121</td>
<td>35</td>
<td>36</td>
<td>53%</td>
<td></td>
<td></td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>Latino(^b) (n=262)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private Insurance (n=217, n=47, n=59)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; High School (n=262, n=43)</td>
<td>99</td>
<td>17</td>
<td>21</td>
<td>31%</td>
<td></td>
<td></td>
<td></td>
<td>0.48</td>
</tr>
<tr>
<td>English Speaking</td>
<td>205</td>
<td>51</td>
<td>67</td>
<td>99%</td>
<td></td>
<td></td>
<td></td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Born in US (n=263)</td>
<td>161</td>
<td>34</td>
<td>55</td>
<td>81%</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Web Q</td>
<td>173</td>
<td>52</td>
<td>66</td>
<td>97%</td>
<td></td>
<td></td>
<td></td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Completed post-education GKS</td>
<td>207</td>
<td>46</td>
<td>65</td>
<td>96%</td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
</tbody>
</table>

Abbreviations: questionnaire (Q), genetic knowledge scale (GKS)

\(^a\) comparison of the three video exposure groups

\(^b\) Non-Latino participants identified as Ashkenazi Jewish

The frequency of users by video exposure time were: (1) minimal (n=269), (2) moderate (n=53), and (3) maximal (n=68). There were a few differences in the demographics by exposure time; participants who were English speaking and completed the study questionnaires online were more likely to have moderate to maximum video exposure time. People who did not complete post-education questionnaire had lower video exposure time (Table 3). Video exposure times were associated with baseline 15 question GKS score (average 11.5 (range 2-15), 11.9 (8-15), 12.6 (6-15), respectively, p-value 0.003).

In the unadjusted analysis there was no difference in the change from baseline to post education in GKS across video exposure groups (Figure 4). After adjusting for the confounders of baseline GKS and insurance status, there remained no association between video exposure time and change in GKS score. In the adjusted analysis, the only predictor of change in GKS score was a lower baseline GKS score (Table 4).

### Table 4. Multiple regression models of three level video exposure time (predictor) with >600 sec. as the reference and primary outcome of change in genetic knowledge scale (GKS) score. Adjusted for confounders association with predictor and outcomes at an alpha < 0.1.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t value</th>
<th>p value</th>
<th>95% CI L</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to &lt;120 sec</td>
<td>-0.32</td>
<td>0.27</td>
<td>-1.17</td>
<td>0.24</td>
<td>-0.85</td>
<td>0.22</td>
</tr>
<tr>
<td>120 to &lt;600 sec</td>
<td>-0.19</td>
<td>0.35</td>
<td>-0.53</td>
<td>0.59</td>
<td>-0.88</td>
<td>0.5</td>
</tr>
<tr>
<td>Baseline GKS</td>
<td>-0.52</td>
<td>0.05</td>
<td>-10.01</td>
<td>&lt;.0001</td>
<td>-0.62</td>
<td>-0.42</td>
</tr>
<tr>
<td>Not Private Insurance</td>
<td>-0.27</td>
<td>0.24</td>
<td>-1.1</td>
<td>0.27</td>
<td>-0.74</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Abbreviations: genetic knowledge scale (GKS), second (sec), standard error (SE) confidence interval (CI), lower (L), upper (U)
Overall, participants rated the educational materials positively with over 85% moderately or strongly agreeing that the website materials helped them understand the different conditions included on the genomic screen, how genomic screening can be helpful to them or their family members and improved their understanding of how genetic variants affect their risk to develop a genetic condition (Figure 5). The overall mean summed score of the website evaluation was 22 with a range of 13-57 on a scale that had a best possible score of 13 and a worst possible score of 84. Participant summed website evaluation score was not associated with video exposure time (data not shown).

**Discussion**

This study is one of the first studies to demonstrate the feasibility and value of conducting detailed participant-specific usage analytics when developing scalable genomic education methods. The collection of log-based participant usage data enabled an exploratory analysis of participants’ characteristics and use patterns, which begins to illustrate the challenges of delivering genomic education through non-traditional, scalable platforms. The majority of participants (83%) expressed interest in learning through the website. Of the 390 participants who requested access, a minority 17% of participants accessed more than 10 minutes of videos. Video usage was not associated with improvement in genetic knowledge as measured by knowledge questions that were part of the study baseline and post-education questionnaires. It is possible that participants accessed other resources through the website besides the videos, for which usage was not directly measured. Additionally,
participants may have already had knowledge of the material included within the videos, and the non-random design of the study allowed participants to self-select which videos to view based upon their perceived baseline knowledge.

The lack of participant usage data for the prior published studies limits our ability to compare our results. As with other studies of alternative educational materials in genomics and other healthcare settings, the majority of the participants who accessed the website endorsed the benefits of the information. Our experience differs from prior studies with regard to a change in participant genomic knowledge and other measurable participant outcomes. These differences are potentially related to a variety of factors though likely a significant factor was the study setting. Our educational tool was used at home and was self-guided, did not have any required content, and the education was not supervised in a clinical or research appointment.

Many of the other studies have administered educational intervention in a healthcare setting or in advance of a clinical appointment. We suspect that their participant usage was higher, because in those settings participants were supervised and given a task to complete and may have felt greater obligation to be compliant as opposed to an unsupervised home setting. However, to be truly scalable, education will need to be delivered in a manner that requires minimal provider supervision and is convenient for the individual. Additionally, our participants who were having genomic screening may have had lower motivation to view the videos compared to participants in other studies who were receiving focused education about a specific indication for which they were referred. Collection of participant usage data across different clinical scenarios will help to better elucidate differences in participant use and influencing factors.

Our experience supports prior concerns that web-based learning is not desirable for all participants and may exacerbate healthcare disparities. Participants who declined access to the website were older, had lower education levels, were more likely to be immigrants, speak Spanish as their primary language and have Medicare or Medicaid than those who chose to access the study website. While it was not measured, they likely had less access to internet devices. These participants as well as participants who requested access to the website but had minimal video usage were much more likely to choose to complete the study questionnaire on paper. In this era of increasing web and electronic based studies, there are individuals who do not have web access and/or are less comfortable learning via a website or video and completing electronic forms. Furthermore, despite efforts to make the website accessible to as many people as possible with translation of content into Spanish, very few people who spoke Spanish as their primary language had meaningful video usage. Other studies have demonstrated effective video health education in minority populations and low literacy populations when the tools have been developed for these specific populations. This suggests one single tool will not meet the needs of all, and it may be necessary to develop multiple tools to address the variable needs of different populations and ensure maximum effectiveness.

The pattern of participant usage provides important insights for future development of self-guided tools. The greatest decline in views occurred in the first five seconds of a video, and participants who watched a video in its entirety were more likely to watch other videos than those who did not. The educational videos developed for this study were visually basic with minimal animation and a voice over. Information was presented in a manner similar to a genetic counseling session; in accessible language and using analogies to explain complex information. This is likely not the most engaging manner in which information could be provided, but our strategy was limited by the research budget. Potentially platforms that immediately engage users with interesting and motivating stories and visually interesting illustrations could capture the viewers’ attention and motivate viewers to continue to watch and increase overall usage.

There are several limitations to our study. Overall, the results of this study are exploratory given the modest sample size. Future, larger, more inclusive studies are needed to confirm our findings. Our study was also restricted to participants who identified themselves as Latino or Ashkenazi Jewish and requested the receipt of genomic screening results; therefore, our findings may reflect unique perspectives of these populations and people who elect to have genomic screening results. Additionally, while the web-platform and video education were available in both English and Spanish languages, only English-speaking participants had meaningful use of the videos. The GKS was an adaptation of other scales and was not validated, and it is possible that there were educational benefits to using the website that were not captured by the GKS. The videos were modified mid-way through the study to reflect changes in the study protocol and make changes to improve the content and presentation; though duplicate analysis of exposure to old or new videos alone did not differ from the analysis of the combined video exposure.

**Conclusion and Practice Implications**

The development of scalable genomic education delivery methods is necessary but is more challenging than anticipated. More user testing with detailed user analytics is needed during the development stage to identify and
adequately address issues of user engagement and retention. Overcoming these hurdles will require collaboration with professionals specialized in the development of interactive websites and video education in order to identify ways to deliver information in an interesting and engaging fashion. Because of the cost associated with robust development, the genomics community needs to collaboratively develop scalable and adaptable tools. Similar to the way genomic sequencing data are shared to advance collective knowledge, existing and emerging health information technologies including effective genomic education methods should be leveraged to enable tailored, layered, and large-scale communication.5

More studies that include detailed user analytics are warranted to continue to explore the effectiveness of novel methods of genomic education in more diverse groups of individuals. Finally, we need to appreciate that even with scalable effective genomic education, there will be individuals for whom traditional models will be preferred and more effective based upon the characteristics of the individual and the clinical context.

Acknowledgments

This research was supported by National Institutes Grants U01HG008680 and UL1TR001873. We would like to thank Aileen Espinal and Bianca Haser for their assistance with participant recruitment. We would also like to thank the participants of the study.

References

Identifying Cancer Patients at Risk for Heart Failure Using Machine Learning Methods

Xi Yang, PhD1, Yan Gong, PhD, MS2,3, Nida Waheed, MD4, Keith March, MD, PhD5, Jiang Bian, PhD1, William R. Hogan, MD, MS1, Yonghui Wu, PhD1

Department of 1Health Outcomes and Biomedical Informatics, 2Pharmacotherapy and Translational Research, 3Center for Pharmacogenomics and Precision Medicine, 4Internal Medicine, and 5Division of Cardiovascular Medicine, College of Medicine, University of Florida, Gainesville, Florida, USA

Abstract

Cardiotoxicity related to cancer therapies has become a serious issue, diminishing cancer treatment outcomes and quality of life. Early detection of cancer patients at risk for cardiotoxicity before cardiotoxic treatments and providing preventive measures are potential solutions to improve cancer patients’ quality of life. This study focuses on predicting the development of heart failure in cancer patients after cancer diagnoses using historical electronic health record (EHR) data. We examined four machine learning algorithms using 143,199 cancer patients from the University of Florida Health (UF Health) Integrated Data Repository (IDR). We identified a total number of 1,958 qualified cases and matched them to 15,488 controls by gender, age, race, and major cancer type. Two feature encoding strategies were compared to encode variables as machine learning features. The gradient boosting (GB) based model achieved the best AUC score of 0.9077 (with a sensitivity of 0.8520 and a specificity of 0.8138), outperforming other machine learning methods. We also looked into the subgroup of cancer patients with exposure to chemotherapy drugs and observed a lower specificity score (0.7089). The experimental results show that machine learning methods are able to capture clinical factors that are known to be associated with heart failure and that it is feasible to use machine learning methods to identify cancer patients at risk for cancer therapy-related heart failure.

Introduction

Cancer is the second leading cause of death in the US.1 There has been a great amount of effort and resources invested in the development of new cancer therapies. The mortality rates of many cancers are being brought under control with the improvement of cancer treatment.2 However, these anticancer treatments often have various side effects. For example, cardiotoxicity is one of the well-documented adverse events of cancer treatments resulting either from accelerated development of cardiovascular diseases in cancer patients or from the direct effects of the treatment on the structure and function of the heart.3 Traditional chemotherapy such as anthracyclines have been known to cause cardiovascular complications.4,6 Cardiotoxicity related to cancer therapies has become a serious issue that diminishes cancer treatment outcomes. A recent study examined various anticancer therapies and reported a significant correlation between quality of life (QoL) and chemotherapy cycles.7 Early detection and possible prevention of cardiotoxicity in cancer treatments is a potential solution to improve cancer patients’ safety and QoL. Identifying cancer patients with high risk of cardiotoxicity is a critical step towards early detection and possible prevention.

In the last two decades, the introduction of targeted anticancer therapies has revolutionized the treatment of both hematological malignancies such as multiple myeloma, chronic myeloid leukemia and solid malignancies such as breast and renal carcinoma.8,9 Contemporary cancer therapy has led to a 23% reduction in cancer-related mortality rate and rapid increase in cancer survivorship in the last 15 years.10 However, some devastating side effects of these treatments have also resulted in increased morbidity and mortality.11,12 Examples of these targeted cancer therapies include human epidermal growth factor 2 inhibitors, inhibitors of vascular endothelial growth factor pathway and tyrosine kinase inhibitors and proteasome inhibitors. Most recently, immune checkpoint inhibitors have also been associated with cardiotoxicity.13,14 Despite the efficacy of these therapies, their widespread use has paradoxically resulted in the emergence of serious cardiovascular effects/complications such as cardiomyopathy/heart failure, coronary artery disease, myocardial ischemia, hypertension, arrhythmia, thromboembolism, and pericardial disease.15 One of the most relevant clinical implications of these complications is treatment interruption, which is associated with cancer recurrence. Due to the high incidence and negative impact on patient outcomes, new medical subspecialties such as Cardio-Oncology were created to optimize the care or management of patients receiving these cancer therapies. Identifying patients with high risk of cardiotoxicity using historical electronic health records (EHRs) could be potentially used to improve cancer treatment safety and QoL.
Rapid adoption of EHRs has made longitudinal clinical data available to research. There is an increasing interest in using longitudinal EHRs to develop computational algorithms for disease onsite prediction. Researchers have applied standard statistical regression models and machine learning methods to predict the onsite of heart failure among general patient cohorts. For example, Wang et al. developed a heart failure predicting model using random forests (RFs) and examined various prediction windows. Sun et al. proposed a method to combine knowledge and data driven method to identify risk factors of heart failure from EHRs. Wu et al. compared three machine learning models including Boosting, support vector machines (SVMs) and logistic regression (LR) for heart failure prediction. While machine learning-based predictive models showed decent performance, previous studies identified issues such as imbalanced data and the lack of modeling temporal sequence among clinical events. Recently, Choi et al. applied recurrent neural networks (RNNs) for heart failure prediction and compared RNN with a traditional machine learning model – SVMs. Their study reported that deep learning models were able to leverage temporal relations among clinical events to improve performance of heart failure prediction with a short observation window of 12-18 months. Rasmy et al. also examined the generalizability of RNN in predicting heart failure onset risk using a large and heterogeneous EHR data set. Researchers from the cancer community have applied statistical regression models for risk assessment of heart failure after cancer treatments. For example, Ezaz et al. applied regression analysis to assess risk scores of heart failure among breast cancer patients after trastuzumab therapy. Authors used a study cohort consisted of women from 67 to 94 years old from the SEER-Medicare database diagnosed with early-stage breast cancer. Although machine learning has been successfully applied to the general patient cohorts for heart failure prediction, it’s not clear whether it can be applied to cancer patient cohorts to support early detection and possible prevention of cardiotoxicity in cardiotoxic cancer therapies.

This study focused on the prediction of cancer patients developing heart failure after cancer diagnoses. We examined four machine learning algorithms for heart failure prediction among cancer patients from the University of Florida Health (UF Health) Integrated Data Repository (IDR). We compared four widely used machine learning models including LR, RFs, SVMs, and Gradient Boosting (GB) for heart failure prediction among cancer patients. We systematically examined variables including patient demographics (gender, race, age), diagnoses, medications, and procedures as machine learning features. We also compared one-hot encoding and term frequency-inverse document frequency (TF-IDF) encoding for all four machine learning methods. Furthermore, we compared the performance of heart failure prediction on general cancer patients with the performance on cancer patients with exposure to chemotherapy drugs. Our ultimate goal is to develop predictive models for identification of cancer patients with high risk of cardiotoxicity to prevent or minimize the risk of cardiotoxicity in cancer treatments.

Methods

Data set

In this study, we used EHR data from UF Health IDR. Supported by the UF Clinical and Translational Institute (CTSI) and the UF Health, the UF Health IDR is a secure, clinical data warehouse (CDW) that aggregates data from the university’s various clinical and administrative information systems, including the Epic electronic medical record (EMR) system. As of February 2019, the IDR contains data for encounters that occurred after June 2011, with a total of more than 1105 million observational facts pertaining to 1.17 million patients. From UF Health IDR, we collected a total number of 143,199 cancer patients from Jan 1st, 2011 to Dec 31st, 2017. We identified cancer patients using the diagnosis codes (at least one cancer diagnosis code). The extracted EHR data contains patient demographics, diagnoses in both the International Classification of Disease version 9 (ICD-9) and the International Classification of Disease version 10 (ICD-10), medications in RXNORM codes, and procedures in CPT codes. This study was approved by the UF Institutional Review Board.

Definition of cases and controls

Starting from the initial cancer patients, we removed patients diagnosed with only benign type of cancers. Then, we identified the first diagnosis date of cancer as cancer index date, denoted as CAID. Following previous studies from Choi et al. and Rasmy et al., we defined the heart failure onset date (HFOD) as the first encounter date of three consecutive heart failure diagnosis encounters occurred within 12 months.

Cases: Following previous studies on heart failure prediction of general patient cohorts, cases of heart failure were defined as follows: (1) having at least three encounters with qualified heart failure diagnoses defined in Table 1 occurred within 12 months; (2) the CAID must before the HFOD – this rule removed the cancer patients with existing heart failure conditions before cancer diagnosis; (3) at least have one medication record or one procedure record – we want to ensure that the patient received treatments at UF Health.
Controls: Controls were defined as cancer patients without any qualified heart failure codes defined in Table 1.

<table>
<thead>
<tr>
<th>ICD-9</th>
<th>Description</th>
<th>ICD-10</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>428</td>
<td>heart failure</td>
<td>I50</td>
<td>heart failure</td>
</tr>
<tr>
<td>428.0</td>
<td>congestive heart failure, unspecified</td>
<td>I50.1</td>
<td>left ventricular failure, unspecified</td>
</tr>
<tr>
<td>428.1</td>
<td>left heart failure</td>
<td>I50.2</td>
<td>systolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.2</td>
<td>systolic heart failure</td>
<td>I50.20</td>
<td>unspecified systolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.20</td>
<td>systolic heart failure, unspecified</td>
<td>I50.21</td>
<td>acute systolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.21</td>
<td>acute systolic heart failure</td>
<td>I50.22</td>
<td>chronic systolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.22</td>
<td>chronic systolic heart failure</td>
<td>I50.23</td>
<td>acute on chronic systolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.23</td>
<td>acute on chronic systolic heart failure</td>
<td>I50.3</td>
<td>diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.3</td>
<td>diastolic heart failure</td>
<td>I50.30</td>
<td>unspecified diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.30</td>
<td>diastolic heart failure, unspecified</td>
<td>I50.31</td>
<td>acute diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.31</td>
<td>acute diastolic heart failure</td>
<td>I50.32</td>
<td>chronic diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.32</td>
<td>chronic diastolic heart failure</td>
<td>I50.33</td>
<td>acute on chronic diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.33</td>
<td>acute on chronic diastolic heart failure</td>
<td>I50.4</td>
<td>unspecified combined systolic (congestive) and diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.4</td>
<td>combined systolic and diastolic heart failure</td>
<td>I50.40</td>
<td>unspecified combined systolic (congestive) and diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.40</td>
<td>combined systolic and diastolic heart failure, unspecified</td>
<td>I50.41</td>
<td>acute combined systolic (congestive) and diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.41</td>
<td>acute combined systolic and diastolic heart failure</td>
<td>I50.42</td>
<td>chronic combined systolic (congestive) and diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.42</td>
<td>chronic combined systolic and diastolic heart failure</td>
<td>I50.43</td>
<td>acute on chronic combined systolic (congestive) and diastolic (congestive) heart failure</td>
</tr>
<tr>
<td>428.43</td>
<td>acute on chronic combined systolic and diastolic heart failure</td>
<td>I50.8</td>
<td>other heart failure</td>
</tr>
<tr>
<td>428.9</td>
<td>heart failure, unspecified</td>
<td>I50.81</td>
<td>right heart failure</td>
</tr>
<tr>
<td>402.01</td>
<td>malignant hypertensive heart disease with heart failure</td>
<td>I50.810</td>
<td>right heart failure unspecified</td>
</tr>
<tr>
<td>402.11</td>
<td>benign hypertensive heart disease with heart failure</td>
<td>I50.811</td>
<td>acute right heart failure</td>
</tr>
<tr>
<td>402.91</td>
<td>unspecified hypertensive heart disease with heart failure</td>
<td>I50.812</td>
<td>chronic right heart failure</td>
</tr>
<tr>
<td>404.01</td>
<td>hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified</td>
<td>I50.813</td>
<td>acute on chronic right heart failure</td>
</tr>
<tr>
<td>404.03</td>
<td>hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage V or end stage renal disease</td>
<td>I50.814</td>
<td>right heart failure due to left heart failure</td>
</tr>
<tr>
<td>404.11</td>
<td>hypertensive heart and chronic kidney disease, benign, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified</td>
<td>I50.82</td>
<td>biventricular heart failure</td>
</tr>
<tr>
<td>404.13</td>
<td>hypertensive heart and chronic kidney disease, benign, with heart failure and with chronic kidney disease stage V or end stage renal disease</td>
<td>I50.83</td>
<td>high output heart failure</td>
</tr>
<tr>
<td>404.91</td>
<td>hypertensive heart and chronic kidney disease, unspecified, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified</td>
<td>I50.84</td>
<td>end stage heart failure</td>
</tr>
<tr>
<td>404.93</td>
<td>hypertensive heart and chronic kidney disease, unspecified, with heart failure and chronic kidney disease stage V or end stage renal disease</td>
<td>I50.89</td>
<td>other heart failure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I50.9</td>
<td>heart failure, unspecified</td>
</tr>
</tbody>
</table>
Data used for prediction: We used all structured EHR data, including patient demographics, diagnoses, medications, and procedures occurred before the HFOD for prediction. Clinical variables occurred at or after HFOD were not used.

Case-control matching: For each case, up to 9 controls were selected according to criteria defined as follows: (1) having the same gender and race as the case; (2) age is within five-year interval of the case; (3) having the same major cancer type; (4) the first encounter date of the control is within a year with the first encounter date of the case; (5) having an encounter occurred either within 30 days before or any time after the HFOD of the case. The encounter date closest to the HFOD is denoted as reference encounter date. For cases that could not be matched to any controls, we removed them from the data set. To facilitate comparison with previous studies, we followed the same case-control matching procedure without replacement. Figure 1 shows an overview of the case-control matching procedure.

Figure 1. An overview of case-control matching procedure

Machine learning algorithms

In this study, we explored four widely used machine learning algorithms including LR, SVMs, RFs, and GB. We applied LR as the baseline method and compared it with SVMs, RFs, and GB. For LR, RFs, and SVMs, we adopted implementations in the scikit-learn library (http://scikit-learn.org). For GB, we used the implementation from the XGBoost package (https://github.com/dmlc/xgboost). For LR, we optimized the optimization method (solver), the regularization parameter $e$, and the tolerance of termination criterion $e$. For SVMs, we used the Radial Basis Function (RBF) kernel and tuned the regularization parameter $e$, and the tolerance of termination criterion $e$. For RFs, we performed the optimization on the parameters including the number of trees (n_estimators), the tree splitting function, the tree max depth (max_depth). For GB, we optimized the learning rate (eta), the maximum depth of a tree (max_depth), and the number of boost trees (n_estimators) and chose the logistic regression for binary classification as the objective function to evaluate the training loss and regularization.

Variable grouping and feature encoding

In this study, we used patient demographics (gender, race, age), diagnoses, medications, and procedures to develop machine learning models for heart failure prediction. As previous studies have identified the sparseness issue of using the diagnoses and medications as features, we grouped the diagnoses codes (ICD-9 or ICD-10) and medication codes (RxNorm concept unique identifier codes or drug names) according to existing disease groupings.

For diagnoses codes, one previous study from Choi et al. grouped the ICD 9 codes using the Clinical Classification Software (CCS) groups (having a total number of 283 groups) and Rasmy et al. grouped ICD 9 codes using PheWAS (Phenome-wide association studies) groups (about 1,800 unique PheWAS groups). Following the study of Rasmy et al., we mapped the ICD-9 and ICD-10 codes to the PheWAS groups. Using the PheWAS code translation table, we mapped a total number of 36,701 unique ICD-9 and ICD-10 codes into a total of 4,973 unique PheWAS groups.

For medication codes in RxNorm concept unique identifiers (RXCUIs), we mapped them to the ingredient-level RXCUIs. For example, Abraxane (with RXCUI 589511) is a brand name for a chemotherapy drug with the ingredient...
Paclitaxel (RXCUI 56946). Thus, we mapped all RXCUI of 589511 to RXCUI 56946 when developing machine learning models. After mapping all medications to their ingredient level RXCUI codes, we mapped a total number of 19,774 clinical-level RXCUI codes to a total of 2,376 ingredient-level RXCUI codes.

To convert the clinical variables into machine-readable features for training, we compared two encoding methods. The first method is one-hot encoding where variables are encoded using binary values of ‘1’ (indicating feature occurred in a sample) or 0 (indicating feature not occurred). The resulted one-hot encoded data is a sparse matrix of 1 and 0. In the one-hot encoding, the frequency of variables is not considered. Variables occurred multiple times were counted only once. In the second encoding method, we further considered the frequency of variables among patients. Inspired by the term frequency-inverse document frequency (TF-IDF) strategy, we treat each patient as a document and each feature as a word in the document. Then, we applied the standard TF-IDF calculation to convert each patient into a TF-IDF weighted vector. For each machine learning algorithm, we trained two models for one-hot encoding and TF-IDF encoding, respectively.

Experiments and evaluation

Using stratified sampling, we split the data into a training set with 13,956 patients (1,566 cases and 12,390 controls) and a test set with 3,490 patients (392 cases and 3,098 controls). We optimized the machine learning models using five-fold cross validation and grid searching. We trained machine learning models using the training set and evaluated the performances using the test set. Following the previous studies, we used the area under the receiver operating characteristic curve (AUC or AUC-ROC) for evaluation. We also conducted statistical tests to compare different methods. The statistical test scores were calculated by sampling 100 times from the test data and each time a number of 1,745 (50% of the test set) samples were randomly selected. We used the t-test to calculate p-values. We also reported the sensitivity and specificity determined using the Youden’s index to facilitate comparison.

Results

Table 2. Comparison of case group and control group.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Sub Types</th>
<th>Case (n=1,958)</th>
<th>Control (n=15,488)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnoses</td>
<td>14.3 (2.1)</td>
<td>12.5 (2.5)</td>
<td>&lt; 0.001*</td>
<td></td>
</tr>
<tr>
<td>Medications</td>
<td>87.7 (47.3)</td>
<td>35.9 (35.4)</td>
<td>&lt; 0.001*</td>
<td></td>
</tr>
<tr>
<td>Procedures</td>
<td>20.4 (5.2)</td>
<td>16.2 (4.3)</td>
<td>&lt; 0.001*</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>1,067 (54.5%)</td>
<td>8,518 (55.0%)</td>
<td>0.691*</td>
</tr>
<tr>
<td>Female</td>
<td>891 (45.5%)</td>
<td>6,970 (45.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>&lt;40</td>
<td>57 (2.9%)</td>
<td>361 (2.3%)</td>
<td>0.007*</td>
</tr>
<tr>
<td>40-44</td>
<td>22 (1.1%)</td>
<td>147 (0.9%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45-49</td>
<td>47 (2.4%)</td>
<td>299 (1.9%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50-54</td>
<td>84 (4.3%)</td>
<td>685 (4.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55-59</td>
<td>169 (8.6%)</td>
<td>1,260 (8.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60-64</td>
<td>197 (10.1%)</td>
<td>1,739 (11.2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-69</td>
<td>263 (13.4%)</td>
<td>2,394 (15.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70-74</td>
<td>335 (17.1%)</td>
<td>2,850 (18.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75-79</td>
<td>275 (14.0%)</td>
<td>2,473 (16.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80-84</td>
<td>249 (12.7%)</td>
<td>1,786 (11.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥85</td>
<td>260 (13.3%)</td>
<td>1,494 (9.6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td>White</td>
<td>1,578 (80.6%)</td>
<td>12,849 (83.0%)</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>African American</td>
<td>304 (15.5%)</td>
<td>1,359 (8.8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>8 (0.4%)</td>
<td>116 (0.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>69 (3.5%)</td>
<td>1,164 (7.5%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*mean value; †standard deviation; ‡derived from the t-test; ‡derived using the chi-squared contingency test
After case-control matching, we identified a total number of 1,958 cases and 15,488 controls. Table 2 compares the descriptive statistics between the case and control groups. For demographic information, the case and control groups have a similar distribution as a result of the matching procedure. On average, patients in the case group have more diagnoses, medications, and procedures compared to the control group patients. In both case and control groups, the most common cancer subgroups include cancer of skin, female breast cancer, prostate cancer, and lung cancer.

Table 3 compares the performance of four machine learning models using two different encoding strategies. The last column shows the statistical p-values between two encoding strategies for each method. The baseline LR achieved an AUC of 0.8795 and 0.8762 for one-hot encoding and TF-IDF encoding, respectively. The statistical test (p-value of 0.1072) showed that there is no significant difference between the two encoding strategies for LR. For other machine learning methods, the TF-IDF encoding outperformed the One-hot encoding with significant p-values. The GB model trained with TF-IDF encoding achieved the best AUC of 0.9077 outperforming all other models with statistical p-values of <0.01. The AUC scores of the SVMs-based models are notably lower than other machine learning models (p-values of < 0.001), which is consistent with the experimental results reported by Wu et al.18 in a similar study. The GB model trained with one-hot encoding obtained the best sensitivity of 0.8546 and the SVMs trained with one-hot encoding obtained the best specificity of 0.8657.

Table 3. Comparison of machine learning methods using one-hot encoding and TF-IDF encoding.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Data Encoding Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>One-hot encoding</td>
<td>0.8418</td>
<td>0.7692</td>
<td>0.8795</td>
<td>0.1072</td>
</tr>
<tr>
<td>LR</td>
<td>TF-IDF encoding</td>
<td>0.7449</td>
<td>0.8589</td>
<td>0.8762</td>
<td></td>
</tr>
<tr>
<td>SVMs</td>
<td>One-hot encoding</td>
<td>0.7423</td>
<td><strong>0.8657</strong></td>
<td>0.8473</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SVMs</td>
<td>TF-IDF encoding</td>
<td>0.7347</td>
<td>0.8518</td>
<td>0.8314</td>
<td></td>
</tr>
<tr>
<td>RFs</td>
<td>One-hot encoding</td>
<td>0.7730</td>
<td>0.8325</td>
<td>0.8737</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>RFs</td>
<td>TF-IDF encoding</td>
<td>0.8240</td>
<td>0.7824</td>
<td>0.8884</td>
<td></td>
</tr>
<tr>
<td>GB</td>
<td>One-hot encoding</td>
<td><strong>0.8546</strong></td>
<td>0.7995</td>
<td>0.8938</td>
<td>0.0016</td>
</tr>
<tr>
<td>GB</td>
<td>TF-IDF encoding</td>
<td>0.8520</td>
<td>0.8138</td>
<td><strong>0.9077</strong></td>
<td></td>
</tr>
</tbody>
</table>

- LR: logistic regression; SVMs: support vector machines; RFs: random forests; GB: gradient boosting; TF-IDF: term frequency-inverse document frequency
- Best AUC, sensitivity, and specificity are highlighted in bold
- p-values were used to compare the performances of the two encoding methods for each machine learning algorithm.

Discussion and Conclusion

In this study, we examined four machine learning algorithms for heart failure prediction using cancer patients’ EHR data from UF Health IDR. Starting from a total number of 143,199 cancer patients, we identified 1,958 qualified cases who developed heart failure after diagnoses of cancers, which were matched to 15,488 controls by gender, age, race, and major cancer type. We compared two feature encoding strategies including one-hot encoding and TF-IDF encoding in developing machine learning methods. The GB model with TF-IDF feature encoding achieved the best AUC score of 0.9077, significantly outperforming other machine learning methods. The experimental results show that it is feasible to use machine learning methods to identify cancer patients with risks of cardiotoxicity. Previous studies have applied machine learning methods for heart failure prediction among general patient cohorts without anchoring on any diseases. This study demonstrated the feasibility of using machine learning methods for heart failure prediction among cancer patients. Among the four machine learning models, LR and RFs achieved decent performance. However, the performance of the SVMs model is relatively lower than other machine learning methods. A similar study by Wu et al.18 also reported that SVMs achieved lower performance in general patient cohorts. We also explored two feature encoding strategies including one-hot encoding and TF-IDF encoding. The experimental results (Table 3) showed that the models trained with the TF-IDF encoding method achieved better performance compared to one-hot encoding method for SVMs, RFs, and GB. A possible reason is that the TF-IDF encoding can capture frequency information of variables thus to enhance performance.

We looked into the subgroup of cancer patients with exposure to chemotherapy drugs to further examine the prediction performance. To facilitate subgroup analysis, two UF Health physicians (NW and KM) manually reviewed a total
number of 1,557 ingredient-level medications used by at least 10 patients and identified 104 chemotherapy drugs. The top three drugs are methotrexate (used by 1,486 patients), cyclophosphamide (used by 1,371 patients), and carboplatin (used by 1,275 patients). Using this list of chemotherapy drugs, we identified 458 patients with exposure to chemotherapy drugs from the test set and calculated the sensitivity and specificity. For this subgroup, the best machine learning model (i.e., GB) achieved a sensitivity of 0.8824 and a specificity of 0.6300, respectively. Compared with the scores on the entire test set (a sensitivity of 0.8520 and a specificity of 0.8138), this subgroup has a better sensitivity but a significantly lower specificity. This may indicate that exposure to chemotherapy drugs is a strong but not deterministic risk factor for heart failure detection. As many chemotherapy drugs are known to have cardiotoxic side effects, it’s not surprising to see that chemotherapy drugs are strong risk factors. Yet, the low specificity indicates that it is still challenging for machine learning models to figure out why some cancer patients develop cardiotoxicity after exposure to chemotherapy treatments while others do not.

This study focused on the prediction of cancer patients who developed heart failure after cancer diagnoses. Compared with the results from general patient cohorts, the machine learning models achieved a better AUC score in our cancer patient cohort constructed from the UF Health IDR. We compared the statistics between our cancer patient cohort with general patient cohorts used in previous studies. The cancer patient cohort used in this study has more rich medication and clinical procedure information. For example, the average numbers of medications and clinical procedures in a general patient cohort constructed by Rasmy et al. from the Cerner Healthfacts® dataset 1 are remarkably lower (27.26 in cases and 1.4 in controls) than our cancer patient cohort (87.7 in cases and 20.4 in controls). This may be one of the reasons that the machine learning methods achieved a better AUC on our cancer patient cohort.

We also calculated the feature importance using the best machine learning model (i.e., GB with TF-IDF encoding method). The results show that diagnosis of essential hypertension, atrial fibrillation, shortness of breath, tobacco use disorder, and hyperlipidemia are the top risk factors for heart failure prediction. These factors are clinically known risk factors for heart failure, indicating that the machine learning models are able to capture important clinical factors for heart failure prediction. Our ultimate goal is to develop predictive models to identify cancer patients with high risks of cardiotoxicity to prevent or minimize the risk of cardiotoxicity in cancer treatments. This study demonstrated the feasibility of machine learning methods for the prediction of cancer patients developing heart failure after cancer diagnoses. Even though machine learning methods can identify important clinical factors, the mechanism of cancer therapy-induced cardiotoxicity is largely unknown. The low specificity score on the subgroup of cancer patients with exposure to chemotherapy drugs shows that it’s still challenging for machine learning models to identify cancer patients with high-risk of cardiotoxicity after chemotherapy. This study is a preliminary step to assess machine learning models for heart failure prediction among cancer patients. To further improve the performance for cancer patients with exposure to chemotherapy drugs, we plan to explore advanced machine learning models such as deep learning and non-clinical factors such as patient genomics data.

This study has limitations. Similar to other EHR-based studies, our study may suffer from incomplete information, varying length of observation, and coding bias. We carefully designed the cancer patients who later developed into heart failure, yet, some of the cases may not necessarily come from cancer treatments. We expect more carefully designed studies to further evaluate our findings.

Acknowledgement

This study was supported by the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences under award number UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Lagged Correlations among Physiological Variables as Indicators of Consciousness in Stroke Patients

Tahsin T. Yavuz¹, Jan Claassen, MD, PhD, FNCS², Samantha Kleinberg, PhD¹
¹Stevens Institute of Technology, Hoboken, NJ; ²Columbia University, New York, NY

Abstract

Consciousness is a highly significant indicator of an ICU patient’s condition but there is still no method to automatically measure it. Instead, time consuming and subjective assessments are used. However, many brain and physiologic variables are measured continuously in neurological ICU, and could be used as indicators for consciousness. Since many biological variables are highly correlated to maintain homeostasis, we examine whether changes in time lags between correlated variables may relate to changes in consciousness. We introduce new methods to identify changes in the time lag of correlations, which better handle noisy multimodal physiological data and fluctuating lags. On neurological ICU data from subarachnoid hemorrhage patients, we find that correlations among variables related to brain physiology or respiration have significantly longer lags in patients with decreased levels of consciousness than in patients with higher levels of consciousness. This suggests that physiological data could potentially be used to automatically assess consciousness.

Introduction

Patients in intensive care units (ICUs) are extremely challenging to treat given the complex, critical, and changing nature of their conditions. Each patient’s state fluctuates over time, due to recovery or worsening, and as a result of interventions. However it can be difficult to gain insight into this state, and to identify factors that can be intervened upon to improve outcomes. That is, clinicians need to understand not only what is happening but why it is happening. One of the most fundamental factors influencing treatment decisions and predictions of prognosis in an ICU is a patient’s state of consciousness. Despite the critical nature of this information, there is not yet a way to measure consciousness automatically. Instead, clinicians administer labor-intensive tests such as the Glasgow Coma Scale (GCS), often only once a day. These subjective daily assessments make it difficult to determine causes and effects of consciousness, which may also fluctuate between assessments. Instead, if consciousness could be measured automatically and continuously, this would have both clinical impact (better understanding of prognosis, earlier identification of changes in consciousness, and reduced burden on clinicians) and research implications (more frequent assessments would enable analysis of causes and effects of consciousness).

While consciousness is still poorly understood, ICUs now routinely collect large amounts of data that can provide insight into a patient’s state. Multimodality monitoring including EEG, cardiovascular parameters, and brain physiology is now being increasingly used to guide clinical care in ICU, and for research predicting ICU outcomes. Despite the success of machine learning in analyzing ICU data, there has not yet been an effort to use routinely collected ICU data to gain insight into consciousness. We focus in particular on automated detection of factors related to consciousness in stroke patients in the neurological intensive care unit (NICU), due to the severity of illness and potential for improving care with a better understanding of consciousness. While data in the NICU is high frequency (often at a resolution of 5 seconds), there are many challenges in using this data to automatically detect consciousness, including the need to combine multiple sensing modalities and identify sometimes subtle signals within noisy data. First, different variables are measured for different patients, necessitating methods that do not require all possible variables to be present. Similarly, we expect different levels of consciousness to manifest indirectly across features in multiple variables. Since consciousness is assessed infrequently (generally once a day), there are few training labels, which poses a challenge for most classification methods. On the other hand, these labels also make it difficult to treat the problem as one of changepoint detection (determining when the system’s parameters or structure change), since we do not know precisely when a change occurred (e.g. if consciousness is high on Monday morning, and low on Tuesday morning we do not know when the change happened in between).

Instead, we propose that multimodality monitoring (MMM) can be used to gain insight into consciousness, by examining lagged correlations between the measured variables. We hypothesize that since many biological variables are
correlated, when consciousness is interrupted, these correlations may be too. Thus, rather than use raw MMM data directly, we propose that the time lag of correlations is a highly informative feature. While prior methods exist to identify lagged correlations, we introduce the first method to identify when the lag between correlated variables changes, and to do so in multimodality data while being robust to noise. We build on cross-correlation, which is widely used to find the lags (delays) in time series, introducing a new approach that can identify when lags change over time. We further contribute a kernel that is better suited to multimodality physiological data (allowing identification of more subtle relationships), and an optimization procedure to identify lags across an entire time series (e.g., patient’s ICU stay) while being robust to noise and outlying values. Experimentally, we show that this method can accurately recover true changing lags on simulated data, while existing work that does not account for changes has significantly lower accuracy. Finally, we apply the approach to a set of NICU data from 61 patients with subarachnoid hemorrhage (SAH), identifying biologically plausible relationships. In particular, we find that brain and cardiovascular variables are correlated with shorter delays in patients with high versus low levels of consciousness.

The primary contributions of our work are:

1. We motivate the use of lagged correlations for gaining insight into biomedical data.
2. We introduce nonstationary normalized cross-correlation ($N^2C^2$), a new method for automatically identifying lagged correlations whose timing may change over time. This highly general approach may be applicable to many types of biomedical time series, as well as time series more generally.
3. We apply $N^2C^2$ to NICU data and demonstrate novel differences in lags that appear to depend on state of consciousness.

Our approach provides a first step toward both better use of high-frequency ICU data to gain insight into previously invisible patient states, as well as a better and more objective understanding of consciousness in the ICU.

Related Work

Consciousness is a fundamental and yet unsolved problem in biology.5, 6 Most studies of consciousness focus on EEG and fMRI data,7 but this does not allow continuous measurement in the ICU. In the ICU, consciousness is generally assessed with qualitative instruments such as the Glasgow Coma Scale (GCS)8 and Alert, Voice, Pain, Unresponsive (AVPU) scale.9 Recent studies have shown that consciousness rests on the brain’s ability to sustain rich brain dynamics10 and metastability in brain dynamics can be used to understand altered levels of consciousness.11 There have been some efforts using bedside EEG to predict changes in consciousness in stroke patients,1 and using network dynamics to distinguish comatose from healthy patients using EEG.12 Other efforts have used EEG data to identify differences in sleep patterns between patients with different states of consciousness.13 However, we are not aware of work using this data to detect the current state of consciousness in real-time.

On the other hand, machine learning analysis of ICU data has focused on predicting readmission14 or ICU outcomes3 and classifying physiological features to identify disorders.15 Yet, physiological data has been shown to provide insight into brain activity, such as in a recent study that identified correlations between physiological signals (including respiration, CO₂ levels in blood, heart rate) and fMRI-monitored brain connectivity.16 Other works showed that delayed cerebral ischemia after SAH could be predicted from physiological time series data.17 Thus, it is possible that physiological data could provide a window into not only clinical outcomes, but also brain function and potentially consciousness. We hypothesize that time lags between variables may be an important signal. For example, lags in resting state fMRI and EEG provided insight into neuronal processes,18 and lags are widely studied among physiological data more generally, including in analysis of electrodermal activity.19 One of the most common methods for examining such lagged relationships is cross-correlation, which has been applied to local field potentials to gain insight into brain regions.20 Cross-correlation is a measure of similarity between two time series. For two time series with identical values, where one is a shifted version of the other, correlation will be highest, with a value of 1. To identify the lag with the maximal correlation for a pair of variables, essentially one time series is iteratively shifted with the correlation being repeatedly computed. Cross-correlation has been used for time delay estimation.21, 22 such
as for finding the phase shift between acoustic signals, but this aims to find the shift between two measurements of the same process, whereas our variables capture different processes that may causally influence one another.

While there are many extensions to cross-correlation, including faster to compute and normalized versions, it still poses problems for biological time series. First, it is assumed that the lag is the same for the entire time series, which is unlikely to be true for a patient whose physiology and state is changing—particularly in the ICU. Second, even during times of relative stability, there are smaller fluctuations due to both physiology as well as the noise and error inherent in biomedical data. By assuming a single lag, these approaches may be susceptible to the effects of outlying values. While some works have applied cross-correlation to nonstationary time series, they have done so by first removing the trends in the data to make it stationary, whereas we propose that the trends are highly informative in ICU and other physiological data where we aim to uncover the causes of changes in state. Most fundamentally, there is currently no method that can identify when and how the lags between variables change over time. We propose that this is a critical signal for assessing consciousness, and methods for finding changes in lagged correlations may have many applications to the other types of health data described above.

Method

As a first step toward automated assessment of consciousness, we focus on identifying how time lags between correlated variables change as a function of consciousness. To do this using physiological data routinely collected in ICU, we introduce a new method, \( N^2C^2 \) (nonstationary normalized cross-correlation), based on normalized cross-correlation (NCC), that identifies when and how the lags between variables changes over time. Unlike the state of the art, which assumes the lag between two time series remains constant, this allows us to identify both gradual shifts in time lag (e.g., slow change from 1 to 5 second lag) as well as sudden transitions (e.g., such as after an intervention). Further, this indirectly provides detection of changepoints, but with more context (i.e., what is changing) and more flexibly (i.e., not requiring a change to be in the mean of a variable’s value), than the state of the art without the need for specifying a model as in many prior works. Thus our output may be that from hour 0 to hour 10 heart rate (HR) and respiration rate (RR) are synchronized with a lag of 0, while from hour 10 to 12, RR lags HR by 1 minute. By testing how the lags differ across patients in different states of consciousness, we can learn characteristics of these states, which could be used for detecting an individual’s current level of consciousness and further when it changes (e.g., when the lag between certain variables grows). Figure 1 shows an overview of the procedure.

We build upon NCC, which is more interpretable than cross-correlation and can handle variables with varying ranges. This is particularly important for MMM data, where variables differ substantially in range, standard deviation, and frequency. NCC is defined as:

\[
NCC(x, y, l) = \frac{\sum_{t \in T} (x[t] - \bar{x})(y[t + l] - \bar{y})}{\sqrt{\sum_{t \in T} (x[t] - \bar{x})^2} \sqrt{\sum_{t} (y[t + l] - \bar{y})^2}}
\]

where \( x \) and \( y \) are time series with the same measurement frequency, \( T \) is the time series, and \( l \) the lag being tested. \( NCC(x, y, l) \) is then the correlation between \( x \) and \( y \) where \( y \) is shifted \( l \) time steps. While NCC can tell us how related two time series are and with what lag, it has the fundamental limitation that the time lag is constant across time.

\( N^2C^2: \) finding changing lags over time

We now introduce nonstationary normalized cross-correlation (\( N^2C^2 \)), which identifies changes in lagged correlations for a set of time series data \( D \). In the case of physiological data, the lags may differ between patients, or within a patient over time (e.g., due to an intervention or change in status). Lags further can fluctuate over time both due to the underlying biology as well as our noisy measurements of it. Thus some seeming changes may be due to relationships that have a time window (e.g., variation within 1-5 min). If we simply accepted the most significant lag at each timepoint, we could then find frequent oscillation between 1 and 5 minutes, even though there is no actual change. To address this we make two key assumptions:

Assumption 1 Lags between physiological variables in ICU tend to zero.
Figure 1: This figure shows an overview of the proposed method. First, we set the length of time window, \( W \). Then, sliding this time window across the series we search over each lag \( l \) in the set of possible lags between \( A \) and \( B \). To identify which lags are significant, we use a permutation test. To ensure lags change smoothly and are not sensitive to outlying values, we determine the lags for the entire time series at once, using the 3D surface shown. Finally, we output a sequence of lags for variables \( A \) and \( B \). In this example, the lags increase non-monotonically across time, with \( B \) having an increasingly delayed response to \( A \)'s value.

Assumption 2 True lags have higher NCC and at a short timescale, lags are stable.

Assumption 1 says that if all else is equal, we believe that shorter lags are more biologically plausible in our application areas. For other applications the reverse may be true, such as when there is a known delay in information transmission. Assumption 2 says that if we find a lag, say, of 3 seconds between HR in a patient’s data at time 0, a lag of 0 at time 1, and lag of 3 seconds for times 2-10, despite assumption 1, the lag at time 1 is unlikely to be zero because we believe lags are locally stable. By stable we mean that having smaller or no changes in lags is more probable than large changes. In practice we may find multiple lags with similar significance levels (perhaps lag 0 was only slightly more significant than 3 at that time), and this is used to ensure smooth transitions at short timescales. While this ensures some smoothness in lags, by examining lags as forming a continuous surface we are still able to identify when there is a change in the trend, such as if lags shift from an average of 0 to an average of 3 over time.

Step 1: Finding potential lags In general, we focus on finding relationships between all pairs of variables \( x, y \) in either direction (with \( x \) lagging \( y \) or vice versa), though this set could be constrained for example when it is known that the correlation must be in one direction or the other. For instance, if we want to find the speed of the sound in a certain material then knowledge of the microphones’ positions can be used to inform the direction of each lag (as it should be from the microphone closest to the source to the others). Note that treating each direction separately (finding the most likely lag from \( x \) to \( y \) and vice versa) enables us to find the lags for each stage of a feedback loop. We assume the data \( D \) is a \( v \times T \) matrix, where \( D(x, i) \) is the value of variable \( x \) at time \( i \). We are able to calculate correlation when some portion of the data are missing, however we also apply methods that have been developed for imputing missing values in ICU and biomedical data to reduce the degree of missingness. More formally, we aim to identify lags that change over time, so we first partition the time series into a set of overlapping windows of size \( w \), with the step size being the measurement frequency (i.e. the window advances by one unit at each step). The parameter \( w \) should be chosen to be small enough that lags are likely to remain stable within this period (so \( w \) should be smaller in data that changes rapidly rather than in systems that are more stable), while being large enough that there are enough measurements to accurately calculate the NCC. While the choice of \( w \) is important, it mainly should be chosen to reflect the scale at which changes are happening, and is not sensitive to small differences. For the ICU data in this paper, we select a \( w \) of 300 minutes (5 hours), since we use data at a resolution of 1 minute. Since there can be many local fluctuations, with a shorter window it becomes more challenging to identify the overall patterns in this noisy data.

Within each window, we iterate over the set of potential lags \( l \in [0, l_{\text{max}}] \), calculating \( NCC(x, y, l) \) using only the data in the window (as again the lag may change over time). That is, for each pair of variables \( x \) and \( y \) in \( D \) for each
window \( w \), we iterate over the lags to form the set of scores:

\[
N^2C^2(x, y, t, w)[t] = \sum_{i \in [t-w, t+w]} (x[i] - \bar{x})(y[i + \ell] - \bar{y}) \sqrt{\sum_{i \in [t-w, t+w]} (x[i] - \bar{x})^2 \sum_{i \in [t-w, t+w]} (y[i + \ell] - \bar{y})^2}.
\]

(2)

The output is for the window centered at time \( t \), a vector of size \( l_{max} \) with correlation scores for each lag. By repeating this for each time \( t \in T \), we have a set of vectors covering the entire time series. Unlike NCC, this uses a subset of the data (rather than the whole time series) at each step, and rather than identify the maximum correlation, outputs the score for all lags.

**Step 2: Assessing significance of lags** As we have stressed, one of the challenges posed by biomedical data is that lags between variables may fluctuate without there being a significant change in the relationship between variables, such as due to noise, measurement error, or relationships with time windows rather than single lags among other reasons. Thus we cannot simply select the lag with the highest correlation at each \( t \), as this may lead to undesirable properties across the time series, namely a violation of our assumption of local stability of lags. To address this, we first identify which correlations are statistically significant, then from that set optimize for a set of smoothly changing lags across the entire time series.

First, not all tested variables will be correlated, so we must determine which pairs (and at which lags) to accept. Since it can be challenging to determine a null hypothesis, we use a permutation test to find the null and then determine which correlations are significant compared to that. The permutation test shuffles the times associated with one variable’s data within the tested window and then recomputes the correlation. The idea is that by breaking the temporal sequence of the data, we can then learn the distribution of scores for time series that are uncorrelated. This is repeated \( S \) times to obtain a distribution of correlation scores for uncorrelated data. In this work we use an \( S \) of 200. Using this, we can determine if, compared to this null distribution, our observed results are statistically significant. The result of this step is a subset of the lags from the first step (those determined to be significant), though there may still be multiple significant lags at each time for each variable pair.

To find the single most likely lag for each pair \( x, y \) at each time \( t \in T \) we apply a kernel to give higher weight to shorter lags (as in our assumptions), and then identify lags using the entire sequence of scores. First, we apply a stretched tricube kernel to \( N^2C^2(x, y, t) \). We use only the positive side of the kernel (\( 0 \leq u \leq 0.5 \)) in:

\[
K(u) = \frac{70}{81}(1 - |u|^3)^3.
\]

(3)

As a result, we have a kernel that encodes a preference for lags closer to zero while still allowing the possibility of longer lags. Figure 2 shows the tricube kernel and our stretched version. As shown, our kernel has the same preference for shorter lags, but decreases more slowly as the lag increases, allowing for identification of longer lags. In this setting, negative lags are captured by testing both \( N^2C^2(x, y, t) \) and \( N^2C^2(y, x, t) \), which further allows identification of feedback relationships.

The kernel encodes assumption 1 (shorter lags are more plausible), but still leaves us with a number of lags at each timepoint. To select between these, we now formulate this as an optimization problem using assumption 2. After applying the kernel, we concatenate each \( N^2C^2(x, y, t) \) vector. This leads to a matrix of size \( T \times l_{max} \), with \( l_{max} \) being the maximum tested lag. Each entry of the matrix is the associated \( N^2C^2 \) score. We can visualize the result as a 3D surface, as shown in figure 1. Intuitively, to select the sequence of lags \([0,T]\) we aim to find the optimal path through the surface. The optimal path is one that maximizes significance while minimizing distance (to avoid large jumps). This optimization problem (max significance, min distance, with a preference for smaller lags as in assumption 2) is akin to voting, and can be solved using the Borda count, applied to sliding windows. Borda count gives a higher ranking to higher correlations and voting among the adjacent lags leads to a stable transition, which fulfills assumption 2. Note that the count is applied to a sliding window (here of length 10, which is sufficiently large to ensure smoothness, while being small enough to allow for changes in lags), which allows for transitions and changes in lags whereas applying it to the entire time series would force consensus. After using Borda count, we have a single series of length \( T \), with the lags between variable \( x \) and \( y \). Note that the series of lags can be discontinuous, since two
variables may not be significantly correlated across the time series. In both physiological data and nonstationary time series (e.g. stock price data), causal relationships may emerge or change across the time series, so two variables that are correlated in one state (regime) may be unrelated in another. Discontinuities may also exist due to missing data.

**Experiments**

We now apply our approach, \( N^2C^2 \), to ICU data from SAH patients to gain insight into consciousness. Before application to the real ICU data, we first evaluate the method on simulated data to demonstrate that it can correctly find changing lags and can do so with higher accuracy than methods that do not account for changing lags.

**Simulated datasets**

Before applying our approach to NICU data, where ground truth is unknown, we first evaluate it on simulated cases. This allows rigorous evaluation in cases where ground truth is known. We aimed to create simulated data that has the relevant properties of our real ICU dataset. While our work aims solely to discover variations in correlation, the structure of our simulation uses causal models to generate observations. We have a set of \( N \) continuous variables, each of which can trigger changes in other variables over a range of lags even when there are no changes (e.g. fluctuation in [1,5] minutes for the entire time series). Other relevant parameters include the probability of activation (i.e. how often does \( y \) change after \( x \) changes) and strength of relationship (i.e. how much does \( x \) change after \( y \) changes). At each timepoint the lag is randomly selected within the actual range, which can also vary over time. Activation probabilities varied from 0 to 1, such as when a relationship is inactive during one time range and true in another.

In this experiment we used \( N = 20 \), generated random connections (allowing feedback loops), set delays as random 10-time unit windows chosen from [1,50] (e.g. [1,10]), chose activation probabilities from Gaussian distribution (\( \mu = 0.8, \sigma = 0.5 \)), set the strength from Gaussian distribution (\( \mu = 0.5, \sigma = 0.5 \)), and used a significance threshold of 0.1 (for consistency with the NICU data). The time lags change 2 to 4 times per relationship across each time series.

We generated 1000 time series with these settings, and evaluate recall (also called true positive rate), which is the fraction of correlations recovered with their true lags (the detected lag is within the actual range for a given time \( t \)). That is, for \( x \) and \( y \) where there is a true correlation between them across time series \( T \), the recall is a summation over each \( t \in T \) with a value of 1 if the identified lag is true, and a 0 if the identified lag is false. Note that this is evaluated for each time in each sequence, creating many opportunities for potentially missed or incorrect lags. Due to the complexity of the structures, and difficulty of evaluating accuracy when there are many complex indirect relationships, we are primarily interested in recall of true relationships. We compare our approach to \( NCC \), which is applied to the entire time series (rather than windows), and uses the same pre- and post-processing otherwise. This comparison allows evaluation of the impact of identifying changing lags rather than a global lag.

Overall, recall for our approach, \( N^2C^2 \), was 78.2%. In contrast, recall for \( NCC \) was 23.4%. Thus we are able to correctly recover the majority of lags at most timepoints for correlated variables, and identify the correlation between
Table 1: Statistically different lagged correlations involving brain-related variables of interest for patients with low and high levels of consciousness. P-value was calculated with the Mann-Whitney U test.

<table>
<thead>
<tr>
<th>Physiological Variables</th>
<th>p-value</th>
<th># of Patients</th>
<th>Mean Lag (Seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>ICP</td>
<td>PbtO2</td>
<td>0.0117</td>
<td>14</td>
</tr>
<tr>
<td>PbtO2</td>
<td>BrT</td>
<td>0.0329</td>
<td>11</td>
</tr>
<tr>
<td>PbtO2</td>
<td>ICP</td>
<td>0.0343</td>
<td>14</td>
</tr>
<tr>
<td>RR</td>
<td>BrT</td>
<td>0.0360</td>
<td>8</td>
</tr>
<tr>
<td>BrT</td>
<td>RR</td>
<td>0.0360</td>
<td>8</td>
</tr>
<tr>
<td>PbtO2</td>
<td>RR</td>
<td>0.0483</td>
<td>9</td>
</tr>
<tr>
<td>RR</td>
<td>PbtO2</td>
<td>0.0631</td>
<td>9</td>
</tr>
<tr>
<td>HR</td>
<td>PbtO2</td>
<td>0.0666</td>
<td>14</td>
</tr>
<tr>
<td>PbtO2</td>
<td>HR</td>
<td>0.0753</td>
<td>14</td>
</tr>
</tbody>
</table>

all correlated variables. However in some cases we identify the correlation with an incorrect time lag. In contrast, NCC finds either the average lag (yet if lags are 10 or 20, the mean lag of 15 will never actually occur) or the most dominant lag (i.e. if one lag is true for 2/3 of the time, that would be identified). Thus we can see it is critical to capture changes in lags. Note that when there are many related causes it can be harder to recover the true timing when examining only correlations. Further, in the generating causal model it is possible to create unfaithful datasets (e.g. with cancelling out), in which case the correlations are impossible to identify, and relationships can be very weak. However, we created these challenging datasets to ensure the approach would be viable in challenging real-world data.

Investigating consciousness in NICU data

We ultimately aim to develop methods to continuously and automatically measure consciousness. Toward this end, we focus on better understanding stroke, using data from subarachnoid hemorrhage (SAH) patients in the NICU at Columbia University Medical Center. We aim to learn whether NCC can identify a signal that can be used to differentiate patients based on their state of consciousness. For each patient we have a set of continuously recorded physiological measurements including respiratory and cardiovascular measures (e.g. respiratory rate, heart rate) and brain-related indicators (e.g. brain oxygenation, brain swelling, blood flow in the brain). However not all variables were measured for all patients and monitors may be started at different times. Values that were missing for 30 minutes or less were imputed using the Fk-NN approach introduced by Rahman et al. and which had high accuracy on this dataset. While variables are measured at varying frequencies, up to every 5 seconds, we use minute-averaged data, as in prior work on discovering causal relationships in this dataset. For all patients, consciousness was assessed by physicians daily during morning rounds between 8 AM and 11 AM. Due to the variation in ICU stay duration, the data contains 302 measurements of consciousness in 61 patients, with a maximum of 18 measurements and minimum of 1 measurement per patient. Consciousness was mapped to a scale from 0 to 5 (introduced in our prior work). In this work we use measurements of 0 (impaired consciousness) and 4 and 5 (intact consciousness, grouped together due to the small number of individuals with 5). We focus on these values due to the challenge of identifying transitions that may occur in between measurements.

As a result, we have 55 assessments of low levels of consciousness (no response) and 88 measurements of high levels (following commands). For each assessment, we examined lagged correlations among all variables in the prior 10 hours, using a sliding window of 5 hours. We tested for lags between all pairs of variables in both directions, unless more than 20% of data in the time window (across both variables) was missing. For significance testing, after the permutation test described in methods (using a one-tailed t-test with p < 0.1 to determine significance) we used the Mann-Whitney U test to determine which lags are significant. We use a threshold of p < 0.1 due to the high variation in data and small sample size.

Table 1 highlights the significant differences across impaired and intact consciousness for brain-related variables. As shown, across all pairs of variables there was a significantly longer delay when patients had lower levels of con-
Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.

Figure 3: Box plots showing differences in time lag of correlations between low and high levels of consciousness. Relationships with $p < 0.1$ are shown. ‘A to B’ indicates that variable B changes after variable A.
Conclusion

A patient’s level of consciousness is a critical piece of information for clinicians determining prognosis and treatment in the ICU. While fully automated detection of consciousness remains a challenging task, we have taken strides toward this end, showing how high frequency ICU data can potentially be used to continuously detect such states. We introduce nonstationary normalized cross-correlation (N$_2$C$_2$) to detect changes in lags between correlated variables. This approach is highly general, and can be used to detect correlations that change over time in many types of time series (e.g. body-worn sensor data, financial time series). We applied our approach to a set of NICU data, and demonstrate how lags between physiological variables differ significantly in SAH patients with high and low levels of consciousness. Overall, variables had significantly lower lags when consciousness was high compared to when it was low. Future work is needed to investigate the exact mechanism behind the differences in lags and whether this feature can be used to continuously detect changes in consciousness. Code is available at: https://github.com/health-al-lab/N2C2

Acknowledgments

This work was supported in part by the NLM of the NIH under Award Number R01LM011826. Thanks to Hadia Hameed, Mark Mirdchouk, and Tara Xu for assistance.

References


Nurses’ Stress Associated with Nursing Activities and Electronic Health Records: Data Triangulation from Continuous Stress Monitoring, Perceived Workload, and a Time Motion Study

Po-Yin Yen, PhD, RN1,2, Nicole Pearl, BS1, Cierra Jethro2, Emily Cooney, BSN, RN2 Brittany McNeil, BS2, Ling Chen, PhD,3 Marcelo Lopetegui, MD, MS4, Thomas M. Maddox, MD5,6, Marilyn Schallom, PhD, RN7

1Institute for Informatics, Washington University School of Medicine, St. Louis, MO, USA; 2Goldfarb School of Nursing, Barnes-Jewish College, BJC HealthCare, St. Louis, MO, USA; 3Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA; 4Instituto de Ciencias e Innovación en Medicina, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, Chile; 5Healthcare Innovation Lab, BJC HealthCare/Washington University School of Medicine, St. Louis, MO, USA; 6Division of Cardiology, Washington University School of Medicine, St. Louis, MO, USA; 7Barnes Jewish Hospital, BJC HealthCare, St. Louis, MO, USA

Abstract

As health IT has become overloaded with patient information, provider burnout and stress has accelerated. Studies have shown that EHR usage leads to heightened cognitive workload for nurses, and increases in cognitive workload can result in stronger feelings of exhaustion and burnout. We conducted a time motion study in an oncology division to examine the relationships between nurses’ perceived workload, stress measured by blood pulse wave (BPw), and their time spent on nursing activities, and to identify stress associated with EHR use. We had a total of 33 observations from 7 nurses. We found that EHR-related stress is associated with nurses’ perceived physical demand and frustration. We also found that nurses’ perceived workload is a strong predictor of nurses’ stress as well as how they spent time with their patients. They also experienced higher perceived mental demand, physical demand, and temporal demand when they were assigned to more patients, regardless of patient acuity. Our study presents a unique data triangulation approach from continuous stress monitoring, perceived workload, and a time motion study.

Introduction

Numerous health information technologies (ITs) have been implemented in hospitals, and influence patient care, nursing activities, and turnover time and volume.1 There is a growing body of literature describing nurses’ work using observational studies.2-10 Several variables have been identified to affect nurses’ activities. For example, patient care time and quality is significantly correlated to nurse staffing and patient acuity.11-14 Patient acuity helps nurse managers determine levels of patient care, staffing, nurse-patient assignment, and the cost of care.15 In addition, the differences in the working pattern among nurses have also been noted.16 Nurses’ perceived workload is significantly correlated to the amount of work that could have been delegated but was not, nurse-patient ratio, and the amount of documentation.17 However, although evidence has shown the impact of these variables, only one study has controlled for these potential confounders to assess nurses’ time allocation on nursing activities.18

In addition, as health IT has become overloaded with patient information, provider burnout and stress has accelerated.19,20 While the amount of literature investigating physician burnout from EHR usage is robust, limited research has explored EHR burnout among nurses. However, studies have shown that EHR usage leads to heightened cognitive workload for nurses, and increases in cognitive workload can result in stronger feelings of exhaustion and burnout.21,22 Similarly, Gregory and colleagues found that nurses’ and physicians’ perception of having insufficient time to complete tasks related to alerts in the EHR was a significant predictor of cognitive weariness, a subscale of burnout in the disseminated questionnaire.23 One study conducted a survey with advanced practice registered nurses exploring the association between EHR related stress and burnout.24 Moreover, more than half of the respondents reported that the EHR added frustration to their day, and insufficient time for documentation in the EHR was commonly reported as well. After adjusting for demographics and practice characteristics, researchers found that EHR frustration and insufficient documentation time were both significant predictors of burnout among nurses. In another survey, 61% of nursing personnel reported feeling heightened frustration when using their respective EHR systems.25 While the
existing evidence discussed suggests that EHR usage leads to burnout among nurses, further investigation must be done in order to properly characterize the association in order to facilitate solutions.

We conducted a time motion study to observe and record nursing activities during their working day shifts between 7am to 7pm. In a prior study, we reported our design and approaches to collect and visualize nursing workflow in three activity dimensions: communication, hands-on tasks and location.\textsuperscript{26,27} Communication represents whom nurses are interacting with; hands-on tasks represent tasks nurses are physically performing (i.e. preparing medication); and location represents where nursing activities take place. We explored these three activity dimensions across the continuum of time to understand multitasking and task switching in nursing practice.\textsuperscript{26,27} We also controlled the distribution of observation time by splitting the 12-hour nursing day shift into three time blocks: 7am-11am, 11am-3pm, and 3pm-7pm. Our novel approach explores the three activity dimensions, across the continuum of time, to understand the phenomena of multitasking in nursing practice.\textsuperscript{26,29-31} The study found that nurses multitasked (having communication and hands-on tasks simultaneously) 39.48\% of all times. The study design and methods also demonstrated a practical and reliable approach to conducting and analyzing time motion studies.\textsuperscript{26,27}

In this paper, we present a new time motion study in an oncology division with the same study design as the previous study,\textsuperscript{26,27} but included additional variables to investigate factors affecting nursing practice, such as staffing, patient acuity, nurses’ perceived workload, and their physiological response to stress, blood pulse wave (BPw). BPw is described as the pulse wave that travels through the circulatory system when the heart contracts and blood is ejected.\textsuperscript{32} BPw is related to cardiac demands,\textsuperscript{32} which represents an elevated or aroused physiological state and is correlated with potential psychological or emotional stressors.\textsuperscript{33} There are few studies about BPw and its relation to stress, but literature regarding pulse wave velocity (PWV) is flourishing. The definition of PWV is comparable to that of BPw, as it is the velocity of the arterial pulse that propagates through the circulatory system.\textsuperscript{33} PWV is used clinically as a measure of arterial stiffness, and is determined by measuring the time it takes a pulse wave to travel a given distance among the blood vessel.\textsuperscript{34,35} A number of studies have discovered associations between mental stress and both PWV and arterial stiffness due to the contribution of stress to increased blood pressure and risk for hypertension.\textsuperscript{36-40} Specifically, acute mental stress has been found to lead to significantly sustained increases in PWV,\textsuperscript{36,37,39} and individuals suffering from chronic stress are more likely to have greater, prolonged PWV values.\textsuperscript{40} The evidence suggests the similarity of BPw and PWV, and that BPw or PWV, can serve as an adequate measure of psychological stress. In this study, we used a commercial device capable of continuously measuring BPw.

The purposes of the study were to examine the relationships between nurses’ perceived workload, stress measured by BPw, and their time spent on nursing activities. We were especially interested in identify stress associated with EHR use. To our knowledge, we were the first study integrating data sources from nurses’ perceived workload, BPw, and a time motion study.

**Methods**

**Setting and Sample**

The time motion study was conducted in an oncology division at a mid-west academic medical center. We recruited registered nurses who worked in the observed unit. We observed nurses in the general patient care and adjacent areas such as the nursing station, hallway, medication/supply room, and patient rooms. In this unit, nurse-patient ratios ranged from 1:3 to 1:4.

**Observed nursing activities**

We observed nursing activities, such as hand-off (shift reporting), direct patient care (patient assessment, medication administration, procedures), indirect patient care (medication preparation, obtaining medication), interprofessional communication, and EHR review and charting. Due to patient safety and the difficulty of observing nurses in patient isolation rooms, we did not document nurses’ hands-on tasks and communication in these rooms. The observable nursing activities list was refined iteratively with the division’s clinical leadership and research team, and finalized during the training and trial observations. In total, we defined 16 types of communication, 35 hands-on tasks, and 23 locations. A list of example activities with definitions and start-end times can be found in our prior publication.\textsuperscript{26}

We used TimeCaT,\textsuperscript{41} a validated electronic time capture tool developed to support data collection for time motion studies, optimized for touch enabled tablet computers and iPads. Observers document nursing activities in TimeCaT when they shadow nurses.\textsuperscript{41} In TimeCaT, we were also able to visualize workflow and explore the location where nursing activities occurred. The data collected in TimeCaT allowed us to portray nurses’ work: with whom the nurse was speaking (communication), while doing what (hands-on task), and at what location.\textsuperscript{26}
Observers & Inter-observer reliability assessment (IORA)

Our observers were three undergraduate senior nursing students. Observers were required to attend training sessions and trial observations for at least 12 hours. We established the inter-observer reliability via the IORA feature provided in TimeCaT.\textsuperscript{41} The IORA in TimeCaT considers four types of agreements: 1) proportion-kappa (P-K): evaluates the naming agreement on virtually created one-second activity, based on the Kappa statistic. P-K provides a global assessment of the agreement over time; 2) naming-kappa (N-K): a systematic pairing approach based on time-overlap, and provides a Kappa statistic representing the agreement on activity naming; 3) duration-concordance correlation coefficient (D-CCC): agreement on the duration of an activity.\textsuperscript{26} We ensured that the IORA results indicated consistent agreement between our observers before the data collection (Table 1).

Table 1. Inter-observer reliability assessment

<table>
<thead>
<tr>
<th></th>
<th>Communication</th>
<th>Hands-On Task</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-K</td>
<td>N-K</td>
<td>D-CCC</td>
</tr>
<tr>
<td>IORA S1 vs S2</td>
<td>93.34%</td>
<td>0.88K</td>
<td>0.905K</td>
</tr>
<tr>
<td>IORA2 S2 vs S3</td>
<td>93.23%</td>
<td>0.86K</td>
<td>0.942p</td>
</tr>
</tbody>
</table>

P-K= proportion-kappa; N-K= naming-kappa; D-CCC= duration-concordance correlation coefficient; S-NW= Sequence-Needleman-Wunsch

Data Collection

No identifiable information or health records were collected. A typical 12-hour nursing day shift was split into three time blocks: 7am-11am, 11am-3pm, and 3pm-7pm. The 4-hour observation time block minimized the chance of unbalanced data if a 12-hour day shift happened to have a heavy or light workload, and also prevented observer fatigue. Each 4-hour observation was a one-on-one observation: one observer shadowed one nurse. Observers maintained a certain distance from the observed nurse during the observation, and were not allowed to interact with the observed nurse in order to collect data reflecting true time duration and context. After approval from the local Institutional Review Board, observers obtained informed consent from the observed nurse as well as permission from patients to observe their care.

Stress: We measured BPw as the indication of the nurse’s stress level.\textsuperscript{32} We used a wearable device on the upper arm of the nurse to continuously measure BPw. The device has been used for monitoring chronic stress and assessment of patients’ mental stress. During the time motion study observations, we provided the wearable BPw measure, developed by Biovotion, to the observed nurse. The wearable device at the upper arm measures biometric signals continuously with timestamps, including BPw. Biovotion has been used for monitoring chronic stress\textsuperscript{45} and assessment of patients’ mental stress.\textsuperscript{44} For each 4-hour observation, the continuous stress (BPw) data were collected and integrated with the time motion study data.

Perceived workload: At the end of each observation, we asked the observed nurse to fill out the NASA task-load index (TLX), which is widely used to assess perceived workload. The NASA-TLX consists of 6 questions:\textsuperscript{44} 1) Mental Demand: How much mental and perceptual activity was required?; 2) Physical Demand: How much physical activity was required?; 3) Temporal Demand: How much time pressure did you feel due to the pace at which the tasks or task elements occurred?; 4) Performance: How successful were you in performing the task?; 5) Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?; and 6) Frustration: How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task? Nurses rated their perceived workload from 0 (very low) to 100 (very high).

Staffing: We collected staffing data for each observation including nurse-patient ratio, patient type (primary, non-primary, or in isolation), number of consecutive days worked, number of discharges by the observed nurse, and patient acuity. We assessed patient acuity based on an adapted AcuityPlus, which is a patient classification methodology developed by the QuadraMed Corporation.\textsuperscript{45} It has been used to determine the acuity and the type of care needed for each patient, such as communication and cognitive support, safety management, activities of daily living, etc. AcuityPlus has been validated by several studies.\textsuperscript{46,47} Based off of the acuity and care, each patient then was assigned a numerical acuity score. The acuity score ranged from 15 (lowest acuity) to 43 (highest acuity).

Nurses’ demographics: We also collected nurses’ age, education, gender, and years of working experience.
Data Analysis

We performed descriptive analysis to summarize nurses’ time spent and stress on nursing activities in communication, hands-on tasks, and locations. We used one-way ANOVA with post hoc analysis and Bonferroni correction to examine group differences, and paired t-test to compare two variables as repeated measures in every observation. We performed the linear mixed effects model to examine factors affecting outcome variables and account for correlation of repeated observations from the same nurse. All statistical analyses were two-sided at significance level 0.05 and conducted in SAS 9.4 and IBM SPSS 25.0.

Results

We completed a total of 33 observations (132 hours) with 7 registered nurses from May 2018 to February 2019. We also conducted a post-hoc analysis and found that our sample size has an 80% power to detect an effect size of 0.6 with linear mixed effects model at significance level 0.05. This was calculated based on a two-sided Wald test of regression coefficients in linear mixed effects model adjusting for design effect and correlation of repeated measures from the same nurse. From the seven nurses, six nurses had three to six observations; one nurse had one observation; 4.41 observations on average. Among the 33 observations, nine were on Monday, 1 on Tuesday, 6 on Wednesday, 5 on Thursday, 9 on Friday, 0 on Saturday, and 3 on Sunday; 11 were 7am-11am, 10 were 11am-3pm, and 12 were 3pm-7pm. Participating nurses have a mean age of 30, with 3.64 years of working experience on average. One male nurse participated, and all nurses have a bachelor’s degree. Table 2 summarizes the demographics, staffing, and nurses’ perceived workload, stress by BPw, and time spent in patient rooms per 4-hour observation.

Before compositing patients’ acuity scores for a single observation (nurse patient ratios ranged from 1:3 to 1:4), we examined whether nurses’ time spent in the patient room was being influenced by patient acuity or patient types: primary patient (P), non-primary patient (NP), and in isolation (ISO). One-way ANOVA showed that there were no differences of nurses’ time spent in a patient room among the three patient types (p=0.8635). Patient acuity was also not a significant predictor for time spent in a patient room (p=0.0847). However, there were differences of the mean patient acuity scores among the three patient types (P: NP: ISO, mean(SD) = 18(1.27): 19.35(1.48): 20.22(2.10), p<0.0001). The patient acuity scores for ISO and NP are significantly higher than P. We used the average acuity score and number of patients as the patient load in an observation. We also computed weighted acuity scores to adjust for the average acuity score in an observation. We weighted patient acuity scores with weights as following: weight_P=18/(18+19.35+20.22); weight_np=19.35 / (18+19.35+20.22); weight_iso=20.22/(18+19.35+20.22). As a result, the single patient acuity score for each observation can be calculated as weighted acuity = weighted acuity total/number of patients. During the analyses, we did not find average acuity or weighted acuity as a significant factor to stress or time spent.

Table 2. Summary of descriptive analysis

<table>
<thead>
<tr>
<th>Age</th>
<th>Years of Working Experience</th>
<th>NASA-TLX</th>
<th>Mental Demand</th>
<th>Physical Demand</th>
<th>Temporal Demand</th>
<th>Performance</th>
<th>Effort</th>
<th>Frustration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>30.00</td>
<td>6.00</td>
<td>3.18</td>
<td>2.38</td>
<td>28.51</td>
<td>48.64</td>
<td>25.31</td>
<td>52.77</td>
</tr>
<tr>
<td>SD</td>
<td>6.00</td>
<td>2.87</td>
<td>0.75</td>
<td>0.70</td>
<td>0.75</td>
<td>48.64</td>
<td>0.75</td>
<td>24.72</td>
</tr>
<tr>
<td>Min</td>
<td>23</td>
<td>0</td>
<td>7.50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max</td>
<td>41</td>
<td>100</td>
<td>7.00</td>
<td>100</td>
<td>95</td>
<td>100</td>
<td>95</td>
<td>95</td>
</tr>
</tbody>
</table>

We examined nurses’ time spent on nursing activities (Table 3) and found that nurses spent the most time communicating with patients (28.07 mins) and other nurses (24.41 mins), followed by phone calls (6.45 mins), and communication with family (6.15 mins). For hands-on tasks, nurses’ spent most time on EHR (45.54 mins), and the time is significantly higher than the second highest hands-on task: direct medication time (15.17 mins) (p=0.000). Nurses spent the most time in patient rooms (84.18 mins) and at the nursing station (80.44 mins).
Factors associated with nurses’ average stress

We performed linear mixed effect models to examine factors associated with nurses’ average stress in a 4-hour observation. When considering the simplest linear mixed effect models with each individual fixed effect, only perceived physical demand, temporal demand, effort, and frustration shows statistically significant (Table 4). We also did not find any difference in stress across time blocks (7am-11am, 11am-3pm, and 3pm-7pm) (p=0.759).

Table 4. Nurses’ average stress with individual fixed effect

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
<th>p value</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental demand</td>
<td>0.04270</td>
<td>0.00215</td>
<td>1.981</td>
<td>0.057</td>
<td>-0.00137</td>
<td>0.00867</td>
<td></td>
</tr>
<tr>
<td>Physical demand</td>
<td>0.06583</td>
<td>0.00239</td>
<td>2.755</td>
<td>0.010</td>
<td>0.001691</td>
<td>0.011475</td>
<td></td>
</tr>
<tr>
<td>Temporal demand</td>
<td>0.05401</td>
<td>0.001965</td>
<td>2.749</td>
<td>0.010</td>
<td>0.001381</td>
<td>0.009421</td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>0.00626</td>
<td>0.002206</td>
<td>2.84</td>
<td>0.010</td>
<td>-0.003887</td>
<td>0.005139</td>
<td></td>
</tr>
<tr>
<td>Effort</td>
<td>0.00563</td>
<td>0.002487</td>
<td>2.273</td>
<td>0.031</td>
<td>0.000568</td>
<td>0.010737</td>
<td></td>
</tr>
<tr>
<td>Frustration</td>
<td>0.00312</td>
<td>0.002111</td>
<td>2.425</td>
<td>0.022</td>
<td>0.000796</td>
<td>0.009446</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.034502</td>
<td>0.017017</td>
<td>2.027</td>
<td>0.010</td>
<td>-0.009901</td>
<td>0.078905</td>
<td></td>
</tr>
<tr>
<td>Years of working experience</td>
<td>0.039901</td>
<td>0.055128</td>
<td>0.724</td>
<td>0.503</td>
<td>-0.103301</td>
<td>0.183104</td>
<td></td>
</tr>
<tr>
<td>Gender (female vs. male)</td>
<td>0.12364</td>
<td>0.363850</td>
<td>0.340</td>
<td>0.748</td>
<td>-0.813758</td>
<td>1.061051</td>
<td></td>
</tr>
<tr>
<td>Number of discharges</td>
<td>0.073314</td>
<td>0.08195</td>
<td>0.861</td>
<td>0.397</td>
<td>-1.01732</td>
<td>2.48360</td>
<td></td>
</tr>
<tr>
<td>Number of patients</td>
<td>0.169840</td>
<td>0.105373</td>
<td>1.612</td>
<td>0.119</td>
<td>-0.046431</td>
<td>0.386112</td>
<td></td>
</tr>
<tr>
<td>Average acuity</td>
<td>0.051950</td>
<td>0.079053</td>
<td>0.657</td>
<td>0.516</td>
<td>-0.110012</td>
<td>0.213911</td>
<td></td>
</tr>
<tr>
<td>Weighted acuity</td>
<td>0.146399</td>
<td>0.174704</td>
<td>0.838</td>
<td>0.409</td>
<td>-0.211087</td>
<td>0.503885</td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable: nurses’ average stress; Fixed effects: single effect.

Nurses’ average stress and perceived workload

When examining multiple fixed effects in the linear mixed effect model to account for demographics and staffing together, none showed significant association with nurses’ stress (Table 5). Also, none of nurses’ demographics and staffing variables were significantly associated with nurses’ stress.

Table 5. Nurses’ average stress and NASA-TLX

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
<th>p value</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental demand</td>
<td>0.002571</td>
<td>0.002340</td>
<td>1.098</td>
<td>0.282</td>
<td>-0.002241</td>
<td>0.007382</td>
<td></td>
</tr>
<tr>
<td>Physical demand</td>
<td>0.04775</td>
<td>0.002694</td>
<td>1.772</td>
<td>0.088</td>
<td>-0.000766</td>
<td>0.010316</td>
<td></td>
</tr>
<tr>
<td>Temporal demand</td>
<td>0.003491</td>
<td>0.002372</td>
<td>1.472</td>
<td>0.153</td>
<td>-0.001384</td>
<td>0.008367</td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>-0.00732</td>
<td>0.002210</td>
<td>3.311</td>
<td>0.137</td>
<td>-0.002583</td>
<td>0.003818</td>
<td></td>
</tr>
<tr>
<td>Effort</td>
<td>0.00502</td>
<td>0.002752</td>
<td>1.273</td>
<td>0.214</td>
<td>-0.002155</td>
<td>0.009159</td>
<td></td>
</tr>
<tr>
<td>Frustration</td>
<td>0.003666</td>
<td>0.002467</td>
<td>1.486</td>
<td>0.150</td>
<td>-0.001423</td>
<td>0.008755</td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable: nurses’ average stress; Fixed effects: single perceived workload, age, years of experience, number of discharges, number of patients, and weighted acuity.

Nurses’ EHR stress and perceived workload

When examining nurses stress while using EHR (EHR charting and review) and their perceived workload, we found that nurses with higher perceived physical demand and frustration were associated with higher stress (p<0.05) (Table 6). But again, none of the nurses’ demographics or staffing showed statistical significance.
Table 6. Nurses’ EHR stress and NASA-TLX

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
<th>p value</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Mental Demand</td>
<td>.003155</td>
<td>.002532</td>
<td>1.246</td>
<td>.224</td>
<td>-.002049</td>
</tr>
<tr>
<td>Physical Demand</td>
<td>.006611</td>
<td>.002825</td>
<td>2.340</td>
<td>.027</td>
<td>.000801</td>
</tr>
<tr>
<td>Temporal Demand</td>
<td>.003246</td>
<td>.002603</td>
<td>1.247</td>
<td>.224</td>
<td>-.002105</td>
</tr>
<tr>
<td>Performance</td>
<td>-.001061</td>
<td>.002409</td>
<td>-1.247</td>
<td>.224</td>
<td>-.000617</td>
</tr>
<tr>
<td>Effort</td>
<td>.004916</td>
<td>.002938</td>
<td>1.673</td>
<td>.106</td>
<td>-.001124</td>
</tr>
<tr>
<td>Frustration</td>
<td>.005592</td>
<td>.002579</td>
<td>2.169</td>
<td>.040</td>
<td>.000271</td>
</tr>
</tbody>
</table>

Dependent variable: nurses’ EHR stress; Fixed effects: single perceived workload, age, years of experience, number of discharges, number of patients, and weighted acuity.

**Nurses’ EHR stress and EHR use time**

When examining nurses’ EHR stress and the time spent on EHR, we did not find nurses’ EHR stress associated with longer use of EHR (p=0.846) (Table 7). None of nurses’ demographics or staffing showed significant different either. We also did not find longer EHR use with more nurse-patient ratio (p=0.733).

Table 7. Nurses’ EHR stress and EHR use time

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
<th>p value</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>EHR time</td>
<td>0.000011</td>
<td>0.000056</td>
<td>.196</td>
<td>.846</td>
<td>-.000106</td>
</tr>
</tbody>
</table>

Dependent variable: nurses’ EHR stress; Fixed effects: EHR time, age, years of experience, number of discharges, number of patients, and weighted acuity.

**Nurses’ time in own patient room and EHR use time**

We examined nurses’ time in own patient room and EHR use time (Table 8). Although not statistically significant, there is a trend that nurses who spent more time using EHR would have less time in own patient room (p=0.092) (Figure 1).

Table 8. Nurses’ time in own patient room and EHR use time

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
<th>p value</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>EHR time</td>
<td>-1.141041</td>
<td>.080317</td>
<td>-1.756</td>
<td>.092</td>
<td>-.307042</td>
</tr>
</tbody>
</table>

Dependent variable: average time in own patient room; Fixed effects: age, years of experience, number of discharges, number of patients, and weighted acuity.

**Stress with own patients vs. other patients**

We also examined nurses’ stress in patient rooms, both paired t-test and Wilcoxon signed ranks test showed that nurses have similar stress in own patients room and in other patients room (own patients room stress: other patients’ rooms stress respective, Mean(SD) = 2.40(0.12): 2.56(0.73), p=0.180).

**Nurses time spent in own patients room and patient acuity**

Besides nurses’ stress, we were interested in whether nurses’ time spent in own patients room being influenced by patient acuity. The result from the linear mixed model, after accounting for nurses’ demographics and staffing, was similar to what we found with one-way ANOVA. We did not find patient acuity associated with nurses’ time spent in own patient’s room (p=0.895), nor any nurses’ demographics and staffing variables (p>0.3).

**Nurses time spent in own patients room and perceived workload**

We also examined whether nurses time spent in patient rooms was associated with their perceived workload. The results showed that nurses’ time spent in own patient rooms had a positive relationship with perceived physical demand, temporal depend, and effort. In other words, when nurses expressed increases in perceived physical demand, temporal demand, and effort, they have likely spent more time in their patient rooms (Table 9).
Several studies have investigated the association between nurses’ workload and the quality of care they provide. One study found that higher workloads, higher acuity assignments, and poorly organized work environments were all key predictors of omitted or significantly delayed patient care. Similarly, another study found that larger patient to nurse ratios contributed to increased missed care. Beyond patient volume and acuity, researchers also found that nurses’ subjective or perceived quality of care and patient safety were strongly related to the number of missed care events. In addition, nurses’ perceived workload contributed to their ability to provide quality care. Using the NASA-TLX and nurse-reported missed essential care reports, researchers determined that even though staffing ratios and patient acuity were strongly associated with missed nursing care, their effects were greatly reduced when taking into account the nurses’ perceived workload. Our findings are consistent with these studies that nurses’ demographics, staffing, and patient acuity were not statistically associated with nurses’ stress. Instead, their perceived workload plays an important role in their time allocation and vice versa. Thus, nursing workload coupled with how that workload is experienced.

### Discussion

**EHR stress associated with nurses’ frustration**

Provider burnout and stress has accelerated due to being overloaded with patient information. Studies have shown that EHR usage leads to heightened cognitive workload for nurses, and increases in cognitive workload can result in stronger feelings of exhaustion and burnout. Similarly, a study found that nurses’ and physicians’ perception of having insufficient time to complete tasks related to alerts in the EHR was a significant predictor of cognitive weariness, a subscale of burnout in the disseminated questionnaire. Additionally a survey of advanced practice registered nurses explored the association between EHR related stress and burnout. More than half of the respondents reported that the EHR added frustration to their day, and insufficient time for documentation in the EHR was commonly reported as well. After adjusting for demographics and practice characteristics, researchers found that EHR frustration and insufficient documentation time were both significant predictors of burnout among nurses. In another survey, 61% of nursing personnel reported feeling heightened frustration when using their respective EHR systems.

In our study, we collected continuous stress data and perceived workload, and coupled with the time motion study, enriched our ability to examine nurses’ EHR stress and validate nurses’ perceived frustration. Our results showed that nurses spent 45.54 minutes per 4-hour time block using the EHR, and this EHR use time was much more than any other communication or hands-on activities. We also found that nurses’ perceived physical demand and frustration were significantly associated with EHR stress. Previous studies reported frustration of using EHR due to insufficient documentation time. It seems ironic that nurses have spent the most time on EHR among all activities, are frustrated, but need more time on EHR. Although we did not detect association between nurses’ stress and longer use of EHR, the insignificant results may be due to that we do not have sufficient power to detect small effects (type II error). We also noticed that the continuous stress monitoring is better at detecting high stress in safety-related events, such as when being the second check witness and in the isolation room. More studies are needed to investigate the use of continuous stress monitoring to detect EHR-related stress. A better EHR design is also required to decrease EHR documentation time and cognitive load in using EHR.

**Nurses’ perceived workload**

Several studies have investigated the association between nurses’ workload and the quality of care they provide. One study found that higher workloads, higher acuity assignments, and poorly organized work environments were all key predictors of omitted or significantly delayed patient care. Similarly, another study found that larger patient to nurse ratios contributed to increased missed care. Beyond patient volume and acuity, researchers also found that nurses’ subjective or perceived quality of care and patient safety were strongly related to the number of missed care events. In addition, nurses’ perceived workload contributed to their ability to provide quality care. Using the NASA-TLX and nurse-reported missed essential care reports, researchers determined that even though staffing ratios and patient acuity were strongly associated with missed nursing care, their effects were greatly reduced when taking into account the nurses’ perceived workload. Our findings are consistent with these studies that nurses’ demographics, staffing, and patient acuity were not statistically associated with nurses’ stress. Instead, their perceived workload plays an important role in their time allocation and vice versa. Thus, nursing workload coupled with how that workload is experienced.

### Table 9. Nurses’ time spent in own patient rooms and NASA-TLX

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
<th>p value</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand</td>
<td>6.438061</td>
<td>3.429993</td>
<td>1.877</td>
<td>0.074</td>
<td>-0.695453</td>
<td>13.571576</td>
<td></td>
</tr>
<tr>
<td>Physical Demand</td>
<td>10.357601</td>
<td>3.801834</td>
<td>2.724</td>
<td>0.015</td>
<td>2.327633</td>
<td>18.387569</td>
<td></td>
</tr>
<tr>
<td>Temporal Demand</td>
<td>9.153769</td>
<td>3.102006</td>
<td>2.951</td>
<td>0.009</td>
<td>2.592586</td>
<td>15.714952</td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>5.682530</td>
<td>3.374887</td>
<td>1.684</td>
<td>0.104</td>
<td>-1.257861</td>
<td>12.62921</td>
<td></td>
</tr>
<tr>
<td>Effort</td>
<td>10.400754</td>
<td>3.369643</td>
<td>3.087</td>
<td>0.010</td>
<td>2.986577</td>
<td>17.814932</td>
<td></td>
</tr>
<tr>
<td>Frustration</td>
<td>7.310985</td>
<td>3.909389</td>
<td>1.870</td>
<td>0.073</td>
<td>-0.725709</td>
<td>15.34768</td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable: nurses’ time spent in own patient rooms; Fixed effects: single perceived workload, age, years of experience, number of discharges, number of patients, and weighted acuity.

**Nurses’ patient load and perceived workload**

We also noticed that nurses with more patients expressed a higher perceived mental demand (p=0.044), physical demand (p=0.041), and temporal demand (p=0.010). In addition, higher perceived frustration is associated with higher patient acuity (p=0.014). All models have accounted for nurses’ demographics and staffing.
may directly affect patient care quality and delivery. This could also be problematic due to the fact that nurses frequently experience interruptions, which was found to be associated with an increased perceived workload.51

Nurses’ time spent on direct patient care and EHR- is there a balance?

Although not statistically significant, we found a trend that nurses’ increased EHR use was associated with less time in own patient room. Also, it seems rational that nurses would spend more time with patients with higher acuity, but we did not find any significant differences in nurses’ time spent based on patient types and acuity. After adjusting for repeated measures, we still did not find any differences due to patient acuity, nurses’ demographics, or staffing variables. Previous research has found that patient care time and quality are significantly correlated to nurse staffing and patient acuity.11-14 However, the complexity and variety of what nurses’ actually do may have impeded nurses’ functions and responsibilities in patient care. Nurses expressed lack of time being with their patients and frustration on insufficient resources needed.32,53 While top-of-license practice is recommended and needed for quality patient care, it remains critical, but unknown, regarding the best strategies to support the health IT world that nurses currently and will continue to face.

Study Limitations

The study had several limitations, 1) one unit in one hospital within one academic health system; 2) The small sample size (n=33); 3) Potential insufficient power to detect small effects (type II error); 4) Utilizing only one commercial device for BPw. Although some studies have utilized the wearable device to assess patients’ stress, we did not test the reliability of the device prior to the study. However, the wearable stress monitoring device did detect sensible stress outcomes. For example, nurses had the highest stress when being the second check witness and in the isolation room; Other limitations include 5) Lack of documented activities in the isolation room, as observers were not allowed access to the isolation room for safety reason; 6) Participation bias as the study required nurses’ participation and consent 7) Observation bias due to human capacity to accurately record every action.54 We may have missed some activities in rapidly changing activities, but we minimized the inconsistency between observers through rigorous training, JORA, and clearly defined definitions of each activity; 8) We detected association between variables, but cannot conclude the causal relationship. Additional qualitative research or a larger sample size would be needed to validate the findings.

Conclusion

We conducted a time motion study to examine the relationship between nurses’ stress and perceived workload, and their association with particular nursing activities. Our study presents a unique data triangulation approach from continuous stress monitoring, perceived workload, and a time motion study, and could be replicable to other healthcare organizations or disciplines. The approach enriched our ability to detect EHR-related stress and was validated by nurses’ perceived physical demand and frustration. Our findings also showed that nurses’ perceived workload is a strong predictor of nurses’ stress as well as how they spent time with their patients. They also experienced higher perceived mental demand, physical demand, and temporal demand when they are assigned to more patients, regardless of patient acuity. Future studies should investigate the use of continuous stress monitor to identify potential support needed, as well as internal stress reduction techniques that could help with reducing perceived stress for nurses.

Acknowledgement

The project was supported by the Healthcare Innovation Lab at BJC HealthCare/Washington University School of Medicine. We especially appreciate the support from the study site, nurse managers, and nurses who participated in the study.

References


Patient Messaging Content Associated with Initiating Hormonal Therapy after a Breast Cancer Diagnosis

Zhijun Yin, PhD, Jeremy L. Warner, MD, MS, Qingxia Chen, PhD,
Bradley A. Malin, PhD
Vanderbilt University, Nashville, TN

Abstract

Hormonal therapy is an effective, but challenging, long-term treatment for patients with hormone-receptor-positive breast cancer. Raising the rate of patients who initiate therapy may be possible by characterizing the factors that influence a patient’s decision. We hypothesized that online patient portal messages convey such factors. To investigate this hypothesis, we focused on breast cancer patients who were prescribed hormonal therapy at Vanderbilt University Medical Center and sent messages through the portal between diagnosis and therapy initiation. We first conducted a topic modeling analysis to generate the main themes of portal messages. We subsequently applied survival analysis to learn the association between the factors conveyed in messages, in term of semantic word groups, and the time elapsed from diagnosis to therapy initiation. We found that consulting with healthcare providers increased the probability of therapy initiation, while mentions of symptoms or negative emotions exhibited a reduced probability.

Introduction

Approximately 12% of U.S. women will develop breast cancer over their lifetime. It is the second most diagnosed cancer (after skin) and has the second highest cancer death rate (after lung) for U.S. women. It is estimated that there will be 268,600 new breast cancer cases in 2019 in the U.S. and over 41,000 women will die from the disease. Hormone-receptor-positive breast cancer makes up 80% of diagnosed cases. In hormone-receptor-positive breast cancer, the cancer cells grow and spread with the assistance of hormones (e.g., estrogen) in the blood. Hormonal therapy, which works by preventing estrogen from stimulating breast cancer cell growth, is an adjuvant (post-surgical) treatment for patients with this type of breast cancer. Evidence suggests that taking hormonal therapy medications, such as tamoxifen, can reduce cancer mortality by one third. As such, it is often recommended that patients take hormonal therapy medications for at least five years.

Adhering to hormonal therapy is not easy for a breast cancer patient. It is reported that nearly 50% of breast cancer patients prescribed hormonal therapy dropped off a regimen before completing a five-year treatment course. There are many factors that may contribute to medication discontinuation behavior. For example, side effects (e.g., depression) can lead to medication discontinuation. As a result, various studies have focused on learning the factors behind why breast cancer patients choose to stop taking a hormonal therapy medication. These studies can be roughly categorized into three classes based on the data that they investigate: 1) interview or survey, 2) structured electronic medical records (EMRs), and 3) user generated content (UGC) in online environments. The first two classes hold merit, but have notable limitations. Generally, studies based on interviews are often time consuming and not scalable to large study cohorts, while survey-based studies are often confined to the pre-defined questionnaires. Studies based on structured EMRs are, on the other hand, limited in that they lack description of treatment experience (e.g., patient’s feelings and emotions). By contrast, UGC has been shown to be an effective resource to learn about a patient’s health related behaviors. For example, studies have shown that the messages that patients send to healthcare providers in a patient portal were indicative of the likelihood of discontinuing hormonal therapy medication.

However, few studies have focused on what factors affect the time at which a breast cancer patient initiates hormonal therapy, relative to their diagnosis. This information is important because it can provide insights into why a patient delays making a decision to start the therapy. While several studies investigated patient decision making, most relied on interviews or qualitative methods. For example, Beryl et al. conducted a longitudinal series of interviews to identify the decision-making process of hormonal therapy. They found that most patients starting a therapy is not a single decision, but rather is a series of decisions. More generally, Marla et al. pointed out that shared decision making needs to center on the person rather than the medical encounter, suggesting the importance of listening to the patient. Thus, in this study, we focused on the secure messages sent by patients to their healthcare providers, one particular type of UGC, in an online patient portal. There are several clinical factors that are likely to affect the decision to start hormonal therapy; e.g., undergoing additional surgery or an unplanned stay in the hospital. We hypothesized that the messages patients convey through online portals contain factors associated with the time from breast cancer diagnosis to hormonal therapy initiation. To investigate this hypothesis, we focused on the EMRs and portal.
communications sent by breast cancer patients prescribed hormonal therapy at Vanderbilt University Medical Center (VUMC). Particularly, we studied patients who sent messages after their diagnosis date, but before taking a hormonal therapy medication. We applied topic modeling to infer the main themes that were discussed in these messages and performed a survival analysis to study the extent to which the themes were associated with the time that breast cancer patients started their treatment.

Methods

Data
This study used de-identified data from the VUMC EMR system and was approved by the Vanderbilt University Institutional Review Board. In this setting, all patient identities were replaced with persistent pseudonyms by a third-party honest broker and all dates within a patient’s records were consistently shifted by a random number of days that were uniformly sampled from the (-365, -1) range. We focused on the patients who were diagnosed with stage I to III breast cancer and were prescribed any of the following hormonal therapy medications: anastrozole, exemestane, letrozole (aromatase inhibitors; AIs), or raloxifene, tamoxifen (selective estrogen receptor modulators; SERMs). We restricted the cohort to those who sent messages between their diagnosis date and the first documentation of a hormonal therapy medication.

Topic Modeling and Word Clustering
Natural language clinical text has high dimensionality, but is also quite sparse. Thus, we to better summarize the content for inference purposes we reduced the dimensionality. There are two general types of methods to realize this goal, both of which we adopted for this project, as they serve different purposes: 1) topic modeling to identify the main messaging themes, 2) word clustering, based on a lower dimensional representation of words (e.g., word embedding terms of word2vec), to create message content predictors for inference.

Topic modeling is a statistical approach for discovering latent topics in a collection of documents or messages. For example, Latent Dirichlet Allocation (LDA) is a generative statistical model that assumes each document can be represented by a small number of topics, where each word in the document can then be generated by one of these topics. The inferred topics can be interpreted by their most relevant terms. Topic modeling is notable because it allows terms to be used in multiple topics and it has often been observed that it groups terms that are similar in their global context. In this study, we applied LDA, as implemented in Mallet Java package (version 2.0.8), to identify the main themes that were communicated in patient portal messages. Since LDA is an unsupervised technique, we adopted the coherence score to determine the optimal number of topics. The coherence score essentially measures the extent to which two high probability terms in a topic appear together in either external documents (e.g., in Wikipedia) or the modeling documents. We selected the number of topics with the highest average coherence score across the proposed topics. This was accomplished by learning LDA models for 2 to 26 topics (with a step size of 1) over all of the messages. To mitigate word sparsity and ensure interpretability, we replaced each term with its lemma form and retained only nouns, verbs, adjectives and adverbs. We also generated bi-gram terms using the genism python package (version 3.6.0) to capture more meaningful phrases. We report on two aspects of this process in our experiments. First, we consider the most salient terms. Saliency of a term is a weighted term frequency that is introduced for better characterizing term importance in describing message topics. The weight is defined as the sum of the Kullback-Leibler divergence between each marginal topic distribution and its conditional distribution on the given term. Second, we consider the topic distribution and sample relevant terms for each topic. The topic distribution is calculated after combining all of the messages into a single document. We display the most terms with the highest distribution in each topic.

Additionally, we applied word clustering to generate semantic word groups for inference purpose. Word clustering relies on a measure to calculate the similarity between two terms. For example, word2vec is a low dimensional representation technique that can be applied to measure semantic word similarity. It represents a group of shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. The premise of word2vec is to apply each term to predict their adjacent neighbors, which represents the word similarity in a local context. If a standard clustering algorithm (e.g., k-means or hierarchical clustering) is applied, a term can only belong to a semantic word group. In this study, we observed that topics inferred from topic modeling techniques, in spite of good interpretability in summarizing the main themes, are poor in inference tasks. This might be due to the fact that the number of inferred topics is very small, which induces substantial information loss. However, increasing the number of topics would increase the possibility of conceptual overlap between topics, which reduces topic interpretability. Directly applying terms expressed in the messages as predictors is a simple method to avoid information loss.
noted earlier, the high dimensionality and sparsity of natural language may not be beneficial for inference. For example, the number of unique terms can be larger than the number of messages. Furthermore, the correlations between terms will require additional regularization in the model, which makes inference more challenging.

In this respect, word clustering can reduce the correlation between term predictors and reduce the dimensionality, which leads to less information loss than topic modeling. To do so, we first trained a word2vec model based on all of the messages sent by breast cancer patients. We then ran an agglomerative hierarchical clustering algorithm with complete linkage, implemented in the scikit-learn python package (version 0.18.1) with 25 to 1000 groups (with a step size of 25). We applied the elbow rule to the standard deviation of the group size, such that we selected the number of groups at the angle where the marginal gain begins to diminish. This was done because, heuristically, a larger group is more likely to contain multiple meanings while a small group is more likely to have little contribution to dimensionality reduction.

Survival Analysis

To perform survival analysis, we investigate how the message content, measured by semantic word groups, is associated with the time from the diagnosis date until a breast cancer patient initiated a hormonal therapy medication.

Statistical Model. We applied a Cox proportional hazards regression model for the survival analysis. The Cox model is applied to investigate the effectiveness of independent predictors with respect to the time when an event of interest occurs. There are two benefits in applying a Cox model in this study. First, the Cox model is semi-parametric and does not assume any particular survival distribution. Second, the Cox model can make use of right-censored patients by incorporating both time (in terms of when an event happened or the latest time without event) and medication use status (e.g., patients who started taking hormonal therapy medications or did not by the end of follow-up) into the model. If the records showed that patients started taking hormonal therapy, then the event occurred. By contrast, if patients did not take any hormonal therapy medication by the end of data observation window, then right censoring occurred. The risk for an independent predictor is represented by the hazard ratio (HR) in the form of the expected exponential of its estimated coefficient in the Cox model. If the HR is bigger (smaller) than 1, then the variable is associated with an increased (decreased) risk of taking a prescribed medication. We used the implementation of the Cox model from the lifelines python package (version 0.9.4) to conduct the survival analysis.

Variables. We defined an observation window of one and a half years from the diagnosis date. The dependent variable was binary, where 1 indicated that a breast cancer patient started taking the medication and 0 indicated that the patient is right censored. The time between the diagnosis date and the start of medication use (or the end of the observation window) was measured in 6-month blocks. The message content variables were constructed as follows: 1) for each patient, we aggregated all the messages sent between the diagnosis date and either the start of medication use or the end of the observation window. As such, we represented each patient as a document of messages; 2) we replaced the words in each message document with their corresponding semantic word group numbers; and 3) we calculated the term frequency – inverse document frequency (TF-IDF) values for each semantic word group in each document, which we subsequently applied as the values of semantic word group variables.

Figure 1. A summary of the patient messages in this study, shown as a histogram of the number of (a) messages sent per patient and (b) terms expressed per message.

Additionally, we introduced four independent variables: 1) the age at diagnosis, 2) race as documented in the EMR, 3) cancer stage at diagnosis, and 4) the number of Current Procedural Terminology (CPT) codes. We dichotomized
the race into Caucasian and non-Caucasian categories. We encoded advanced cancer stage (stage III) with 1 and early cancer stages (stage I and II) with 0. We included the number of CPT codes because, intuitively, the more procedures a patient is affiliated with, the greater the likelihood that they may delay the start of a hormonal therapy regimen. We report the statistically significant features at the 0.05 significance level.

We also applied the Kaplan–Meier (KM) estimator to estimate how the probability of starting hormonal therapy changes over time. In clinical studies, this estimator is often relied upon to measure the proportion of patients who live for a certain amount of time after treatment. We use the implementation of KM in the *lifeline* python package (version 0.9.4) to calculate and visualize the proportion of breast cancer patients who begin hormonal therapy along the treatment timeline.

**Results**

**Cohort Summary**

The study cohort consists of 336 breast cancer patients prescribed hormonal therapy medications between the years 2005 to 2017, which consisted of 30% of all the breast cancer patients with early stages. The average patient age is 54.8 (with a range of [22, 88]). Based on race in the EMR, 90.1% of the cohort is Caucasian (303). With respect to the disease, 89.3% of the cohort were diagnosed with early stage breast cancer (300). The patients generated a total of 3,329 messages. Figure 1(a) depicts the frequency distribution of the number of messages generated per patient. It can be seen that most patients posted a limited number of messages. For example, 71.7% of the patients sent less than 10 messages during this period. By contrast, only 11.7% of the patients sent at least 20 messages. Figure 1(b) summarizes the frequency distribution of the distinct terms per message. It can be seen that most of the messages contained only a small number of terms. For example, 58.9% of the messages contained no more than 50 terms. By contrast, only 12.0% of messages contained at least 100 terms.

**Top Salient Terms**

![Top Salient Terms](image)

*Figure 2.* The 30 most salient terms in the patient portal messages. The terms are ordered in a descending rank according to their saliency.

Figure 2 demonstrates the 30 most salient terms (from top to bottom) in the portal messages. In most cases, the more frequent the word, the higher the saliency; however, there are some exceptions. For example, the terms *time, week, work, question, start* and *leave* have relatively higher frequency but lower saliency, suggesting saliency might be more effective in describing the importance of a term. Figure 2 further illustrates that *surgery, appointment* and *chemo* are among the most salient words, while other terms such as *back pain, result,* and *prescription* have relatively lower saliency.
Message Topics

Table 1. The most relevant terms for each of the six inferred topics. The topic distribution is shown in the rightmost column. The terms are ordered based on their probabilities in each topic in a descending manner.

<table>
<thead>
<tr>
<th>#Topic</th>
<th>Top Most Relevant Terms</th>
<th>Dist.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>feel, pain, leave, night, med, bad, arm, area, lot, continue, normal, sore, sleep, issue, friday, skin, nausea, red, side, fever, hand, infection, hurt, cough, run, swell, taxol, eat, head, chest</td>
<td>17.46%</td>
</tr>
<tr>
<td>2</td>
<td>chemo, breast, time, treatment, week, cancer, start, find, biopsy, infusion, report, year, ye, thing, center, put, chemotherapy, month, recommend, give, begin, mention, read, show, yesterday, drug, post, question, end</td>
<td>16.87%</td>
</tr>
<tr>
<td>3</td>
<td>surgery, appointment, dr, schedule, question, week, time, follow, dateth, radiation, wait, change, visit, reconstruction, plan, long, meet, remove, understand, discuss, make, mastectomy, set, decide, procedure, cancel, thought, lumpectomy, reschedule, lymph_node</td>
<td>16.72%</td>
</tr>
<tr>
<td>4</td>
<td>day, good, morning, back, result, test, blood, hope, great, lab, hour, concern, pm, mg, advise, tomorrow, problem, stop, foot, drain, forget, dear, low, tablet, blood_pressure, daily, weekend, clinic, guess, level</td>
<td>16.44%</td>
</tr>
<tr>
<td>5</td>
<td>phone, work, send, fax, office, prescription, receive, doctor, place, number, pharmacy, information, refill, email, institution, form, return, fmia, order, request, fill, pick, insurance, leave, medicine, hospital, medication, medical, update, letter</td>
<td>16.33%</td>
</tr>
<tr>
<td>6</td>
<td>call, today, appt, make, check, give, yesterday, home, message, thursday, contact, work, port, friday, talk, monday, tuesday, speak, regard, cell, care, place, nurse, husband, hear, tomorrow, vanderbilt, mom, afternoon, move</td>
<td>16.18%</td>
</tr>
</tbody>
</table>

Table 1 shows the six topics that were learned from the messages. Here, we take a moment to provide some intuition into the topics. First, the most distributed (or mentioned) topic (#1) is about symptoms (17.46%). Second, the next most distributed topic (#2) is about chemotherapy (16.87%). Third, the next most distributed topic (#3) is about surgery. Fourth, the next most distributed topic (#4) is about laboratory tests. Finally, and the last two topics (#5, 16.33% and #6, 16.18%) are about prescriptions and communications, respectively. Note that the topic distribution is calculated after combining all messages into a single document.

Model Fitting

Our approach finally generated 150 semantic word groups, which served as message content predictors in the survival analysis. Our Cox model presented a concordance of 0.872, suggesting a good fit. With respect to the four control variables, it was found that the advanced cancer stage indicated a reduced probability of patients initiating hormonal therapy (HR = 0.851, P = 0.001). This suggests that patients with early stage cancer are more likely to begin hormonal therapy. It was also found that the number of CPT codes correlated with a decreased probability of starting hormonal therapy (HR = 0.983, P < 0.001). Neither the age at diagnosis nor the race were found to be statistically significant.

![Figure 3](image_url)

**Figure 3.** The KM curve for the probability that patients did not initiate hormonal therapy. The intersection of the dashed lines indicates that 50% of patients began hormonal therapy 3 months after diagnosis.
Figure 3 shows the KM curve for the probability of not starting hormonal therapy as a function of time. The horizontal dashed line indicates where the probability is 0.5. The dashed lines illustrate that 50% of the patients initiated hormonal therapy at 3.24 months (0.54 × 6-month unit). It can further be seen that approximately 80% of breast cancer patients began hormonal therapy within one and a half years of the diagnosis.

**Semantic Word Groups**

The semantic word groups that are positively associated with hormonal therapy initiation. The terms are ordered based on their similarity with the center of the group in a descending manner.

<table>
<thead>
<tr>
<th>Group #</th>
<th>Word Samples</th>
<th>HR</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>understand, think, suggest, say, recall, agree, remember, know, decide, realize, knew, consider, assume, recommend, believe, guess, choose, forget, thought, determine</td>
<td>1.964</td>
<td>(1.537, 2.510)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>40</td>
<td>purple, yellowish, bumpy, whelps, raw, thick, electrodes, phlegm, puffy, colored, raised, weeping, scalp, scabs, mucous, corners, runny, peeling, pulsing, bruised</td>
<td>1.928</td>
<td>(1.542, 2.410)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>131</td>
<td>upcoming, reconstructive, cataract, pre, post, rescheduling, prior, canceled, scheduling</td>
<td>1.764</td>
<td>(1.444, 2.154)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>59</td>
<td>later, earlier, sooner, maybe, possibly</td>
<td>1.599</td>
<td>(1.328, 1.924)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>106</td>
<td>sugar, glucose, sugars, systolic, creatinine, pcv, ferritin, count, counts, soaked, platelets, consistently, pressures, diastolic, rate, heart rate, dried, values, 101, 99</td>
<td>1.457</td>
<td>(1.179, 1.800)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>24</td>
<td>ideas, insights, suggestions, cancellations, openings, word, idea, cure, conclusions</td>
<td>1.408</td>
<td>(1.162, 1.707)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>39</td>
<td>meet, speak, confer, consult, communicate, talk, discuss, share, speaking, consulting</td>
<td>1.405</td>
<td>(1.152, 1.712)</td>
<td>0.001</td>
</tr>
<tr>
<td>100</td>
<td>baptist, gateway, institution, centennial, vandy, summit, vu, oho, premier, dayani, vumc, regional, umc, vanderbilt, ctr, center, main, campus, location, 100</td>
<td>1.391</td>
<td>(1.126, 1.718)</td>
<td>0.002</td>
</tr>
<tr>
<td>37</td>
<td>inconvenience, inconvenience, confusion, missed, notice, missing</td>
<td>1.379</td>
<td>(1.100, 1.730)</td>
<td>0.005</td>
</tr>
<tr>
<td>99</td>
<td>confirm, clarify, verify, remind, inform, explain, notify, cancel, reschedule, tell, ask, make, schedule, change, mention, postpone, switch, book, miss, attend</td>
<td>1.370</td>
<td>(1.131, 1.659)</td>
<td>0.001</td>
</tr>
<tr>
<td>79</td>
<td>throwing, backed, picking, threw, stirred, screwed, messed, flared, built, lit, clears, picks, woken, speed, build, lined, chaled, clearing, messes, stood</td>
<td>1.360</td>
<td>(1.109, 1.668)</td>
<td>0.003</td>
</tr>
<tr>
<td>120</td>
<td>papers, documents, paperwork, forms, form, paper, release, ticket, parking, intermittent</td>
<td>1.336</td>
<td>(1.079, 1.654)</td>
<td>0.008</td>
</tr>
<tr>
<td>32</td>
<td>hcl, 1mg, 100mg, 50mg, 500mg, 150mg, 40mg, 25mg, 200mg, 4mg, 2mg, 60mg, amlodipine, 75mg, qd, 80mg, bid, tid, 30mg, glipizide</td>
<td>1.328</td>
<td>(1.072, 1.639)</td>
<td>0.009</td>
</tr>
<tr>
<td>21</td>
<td>important, beneficial, safe, alright, confusing, acceptable, frustrating, true, effective</td>
<td>1.306</td>
<td>(1.065, 1.602)</td>
<td>0.010</td>
</tr>
<tr>
<td>27</td>
<td>discontinued, steriods, cipro, flagyl, decadron, stopped, initially, previously, finished</td>
<td>1.297</td>
<td>(1.050, 1.602)</td>
<td>0.016</td>
</tr>
<tr>
<td>135</td>
<td>whatsoever, injuries, masses, avail, worries, evidence, friction, complaints, apparent, ink, hurry, regurgitation, signs, luck, improvement, complications, success, rush, longer</td>
<td>1.279</td>
<td>(1.062, 1.540)</td>
<td>0.009</td>
</tr>
<tr>
<td>3</td>
<td>sunday, saturday, evening, tonight, morning, afternoon, today, yesterday, fell, fall</td>
<td>1.243</td>
<td>(1.014, 1.523)</td>
<td>0.036</td>
</tr>
<tr>
<td>138</td>
<td>avastin, initiation, recon, inhibitors, cytoxan, goserelin, radiation, laparoscopic, stereotactic, tc, minimally, taxotere, shrink, cosmetic, perjeta, subsequent, 2mm, invasive, carpal, navelbine</td>
<td>1.238</td>
<td>(1.026, 1.495)</td>
<td>0.026</td>
</tr>
<tr>
<td>57</td>
<td>sleeve, compression, lymphedema, glove, garment, gauntlet, garments, massage, lymphedema, bra</td>
<td>1.232</td>
<td>(1.016, 1.494)</td>
<td>0.033</td>
</tr>
<tr>
<td>140</td>
<td>hospital, room, rehab, clinic, emergency, er</td>
<td>1.224</td>
<td>(1.009, 1.486)</td>
<td>0.040</td>
</tr>
<tr>
<td>36</td>
<td>handle, figure, imagine, find, tolerate, manage, afford, pass, watch, run</td>
<td>1.215</td>
<td>(1.012, 1.458)</td>
<td>0.036</td>
</tr>
</tbody>
</table>

Table 2 shows the semantic word groups that are positively associated with the start of hormonal therapy. Based on this result, there appear to be four types:
- Mentions of semantic word groups related to cognitive processes (#12, HR = 1.964, P < 0.001), suggestion-related nouns (#24, HR = 1.408, P < 0.001), and communication or consulting related verbs (#39, HR = 1.405 P = 0.001).

- Mentions of semantic word groups related to schedules (#131, HR = 1.764, P < 0.001; #99, HR = 1.370, P = 0.001; #37, HR = 1.379 P = 0.005), locations (#100, HR = 1.391, P = 0.002; #140, HR = 1.224, P = 0.040), and paperwork (#120, HR = 1.336, P = 0.008).

- Mentions of semantic word groups related to symptoms (#40, HR = 1.928, P < 0.001), vital and test related words (#106, HR = 1.457, P < 0.001), medication dosage (#32, HR = 1.326, P = 0.009), discontinuation of certain medications (#27, HR = 1.297, P = 0.016), medications for chemotherapy and radiation (#138, HR = 1.238, P = 0.026), and verbs related to manage health conditions (#36, HR = 1.215, P = 0.001).

- Other semantic word groups, such as time related words (#59, HR = 1.599, P < 0.001; #3, HR = 1.243, P = 0.026), and adjectives (#21, HR = 1.306, P = 0.010).

**Table 3.** The semantic word groups that are negatively associated with hormonal therapy initiation. The terms are ordered based on their similarity with the center of the group in a descending order.

<table>
<thead>
<tr>
<th>Group #</th>
<th>Word Samples</th>
<th>HR</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>watery, soft, irritable, excessive, sweating, caffeine, loose, feverish, wheezing, breaths, achy, shaking, soaking, stools, dry, fullness, severely, sensitivity, heartbeat</td>
<td>0.596</td>
<td>(0.466, 0.761)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>82</td>
<td>difficult, ill, hard, sick, quickly, easy, busy, long, much, rough</td>
<td>0.631</td>
<td>(0.516, 0.771)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>26</td>
<td>referral, referral, favor, personally, explaining, answered</td>
<td>0.635</td>
<td>(0.528, 0.763)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>149</td>
<td>blood, blood, lab, hydrocephalus, dental, excuse</td>
<td>0.645</td>
<td>(0.502, 0.830)</td>
<td>0.001</td>
</tr>
<tr>
<td>33</td>
<td>group, hospice, private, provider, services, program, ymca, coordinator, facility, assisted, department, living, abc, critical, tenn, tri, university, lives, fitness</td>
<td>0.647</td>
<td>(0.504, 0.831)</td>
<td>0.001</td>
</tr>
<tr>
<td>35</td>
<td>abundance, estimate, invoice, ophthalmologist, oversight, official, integral, itemized, eco, exception, appointment, appeal, indigent, mirror, extension, aflac, expert, eob, endo, error</td>
<td>0.652</td>
<td>(0.533, 0.798)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>98</td>
<td>cervical, grafting, sparing, areola, dissection, silicone, radiated, lateral, prolapse, scar, tissue, lumbar, ovarian, surrounding, lat, reduction, implant, partial, uterine, expanders</td>
<td>0.659</td>
<td>(0.537, 0.810)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>10</td>
<td>mile, bed, class, miles, chair, hrs, hour, car, couch, hours</td>
<td>0.668</td>
<td>(0.540, 0.827)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>17</td>
<td>dexe, cat, mug, ct, pet, echo, ekg, xray, density, eng, marrow, pap, dye, bone, contrast, echocardiogram, smear, thyrogen, tail, hida</td>
<td>0.696</td>
<td>(0.560, 0.866)</td>
<td>0.001</td>
</tr>
<tr>
<td>29</td>
<td>thigh, soreness, calf, upper, elbow, numbness, tingling, tenderness, sensation, stiffness, wrist, forearm, abdomen, ankle, weakness, neck, thumb, tightness, shoulder</td>
<td>0.721</td>
<td>(0.581, 0.895)</td>
<td>0.003</td>
</tr>
<tr>
<td>123</td>
<td>reconstruction, mastectomy, lumpectomy, diep, flap, bilateral, procedure, hysterectomy, surgery, surgical, surgeries, mammograms, ovaries, expansion, mris, port, operation, uterus, final, cath</td>
<td>0.73</td>
<td>(0.595, 0.894)</td>
<td>0.002</td>
</tr>
<tr>
<td>126</td>
<td>phone, telephone, number, cell, address</td>
<td>0.737</td>
<td>(0.606, 0.896)</td>
<td>0.002</td>
</tr>
<tr>
<td>81</td>
<td>wks, seconds, nights, days, weeks, rounds, months, sessions, occasions, years</td>
<td>0.748</td>
<td>(0.605, 0.925)</td>
<td>0.007</td>
</tr>
<tr>
<td>133</td>
<td>expected, involved, likely, present, normal, open, available</td>
<td>0.753</td>
<td>(0.615, 0.921)</td>
<td>0.006</td>
</tr>
<tr>
<td>103</td>
<td>2nd, 1st, 3rd, third, second, preop, tentatively</td>
<td>0.755</td>
<td>(0.620, 0.918)</td>
<td>0.005</td>
</tr>
<tr>
<td>20</td>
<td>gyn, urologist, cardiologist, neurologist, dermatologist, gynecologist, oncologist, specialist, ob, neuro, obgyn, onc, rheumatologist, surgeon, doc, internist, psychiatrist</td>
<td>0.769</td>
<td>(0.628, 0.943)</td>
<td>0.012</td>
</tr>
<tr>
<td>89</td>
<td>affects, effect, effects, right, left</td>
<td>0.779</td>
<td>(0.622, 0.976)</td>
<td>0.030</td>
</tr>
<tr>
<td>92</td>
<td>single, ie, multiple, plus, various, including, entire, numerous, whole, throughout</td>
<td>0.788</td>
<td>(0.653, 0.950)</td>
<td>0.013</td>
</tr>
<tr>
<td>41</td>
<td>contacting, emailing, asking, writing, calling, hearing, posted, waiting, hear, heard</td>
<td>0.796</td>
<td>(0.653, 0.971)</td>
<td>0.024</td>
</tr>
<tr>
<td>130</td>
<td>bruises, fingernails, cases, parts, instances, areas, flagged, toenails, spots, organs, bumps, lumps, blockages, sites, separated, cancers, places, layers, factors, situations</td>
<td>0.797</td>
<td>(0.641, 0.992)</td>
<td>0.042</td>
</tr>
<tr>
<td>60</td>
<td>need, want, needed, necessary, wants, possible, plan, decided</td>
<td>0.803</td>
<td>(0.646, 0.998)</td>
<td>0.048</td>
</tr>
</tbody>
</table>
Table 3 shows the semantic word groups that are negatively associated with the start of hormonal therapy. Based on this result, there appear to be three types:

- Mentions of symptoms (#65, HR = 0.596, P < 0.001; #82, HR = 0.631, P < 0.001; #29, HR = 0.721, P = 0.003; #149, HR=0.645, P=0.001), different physician roles (#20, HR = 0.769, P = 0.012), reconstruction and other surgery (#123, HR=0.730, P=0.002), body components (#130, HR = 0.797, P = 0.042), negative emotions (#52, HR=0.811, P=0.033), x-ray related examinations (#17, HR = 0.696, P = 0.001).
- Mentions of fitness service (#33, HR = 0.647, P = 0.001), distances and hours (#10, HR=0.668, P<0.001; #81, HR=0.748, P=0.007) have decreased probability of taking hormonal therapy.
- Mentions of communication related terms (#126, HR = 0.737, P = 0.002; #41, HR = 0.796, P = 0.024).

Discussion and Conclusion

Primary Findings

This study investigated the messages sent by patients through the VUMC patient portal after their breast cancer diagnosis and before the start of hormonal therapy. We characterized how certain factors mentioned by the patients correlated with when they initiated a regimen of hormonal therapy medications. There were several notable findings. First, our exploratory analysis showed most patients sent only a limited number of messages. This is a similar observation to those made about individuals who published posts in online health communities. Second, our topic analysis suggested that patients were primarily communicating about their symptoms, chemotherapy, surgery, laboratory tests and prescriptions. This has face validity because surgery and chemotherapy are two of the most common interventions that take place before hormonal therapy is prescribed. These two interventions often generate side effects that require additional medications to manage. Third, our survival analysis suggested that breast cancer patients with advanced cancer stages and a larger number of procedures (as documented by CPT codes) were associated with a decreased probability of hormonal therapy initiation. This confirms our expectation that CPT codes should be control variables in the survival analysis. The KM curve indicated that most breast cancer patients who initiated hormonal therapy did so within 6 months after their diagnosis.

There are several notable findings from a comparison of the semantic word groups that increased and decreased the probability of hormonal therapy medication initiation. It should be recognized that it is the factors that are mentioned in the messages (e.g., the act of taking surgery), instead of the mentions themselves, that are material to this investigation. First, we observed that patients who mentioned consulting-related words and chemotherapy were more likely to initiate hormonal therapy. This suggests that obtaining suggestions from healthcare providers may help realize a smooth transition from diagnosis to taking hormonal therapy. By contrast, patients who mentioned symptoms and surgeries were less likely to start hormonal therapy, suggesting that complex health conditions or significant procedures (e.g., surgery) may delay the start of taking hormonal therapy. In particular, women who must undergo repeated breast conserving surgeries due to positive margins, as well as those who develop post-surgical complications (e.g., infection), may be less likely to initiate hormonal therapy; this phenomenon has been studied in the context of adjuvant radiotherapy. It should be noted that, while symptoms-related semantic groups were found to be both positively and negatively associated with the start of hormonal therapy, they are referring the different symptoms (e.g., #40 against #29). Second, it was observed that patients who mentioned negative emotions (e.g., depression and anxiety) were less likely to initiate the therapy. By contrast, patients who mentioned afford related verbs (#36) or acceptable related adjectives (#21) were more likely to start therapy. This emphasizes the importance of complex health conditions in affecting the start of therapy. Finally, we found that patients who mentioned scheduling related terms are more likely to start therapy. This suggested that actively managing appointments with healthcare providers is beneficial to start therapy.

Limitations and Future Work

Despite these findings, we wish to highlight several limitations, which serve as the basis of future research. First, our findings were generated from a cohort that were predominantly Caucasian and diagnosed with early stage breast...
cancer at a single medical center, which may limit their generality. Second, we are unable to discern whether hormonal therapy was prescribed in the adjuvant or neo-adjuvant (pre-surgical) setting. Neoadjuvant hormonal therapy was intensively studied in the early 2000’s but then fell out of favor; recently it has been used more often in the context of “window of opportunity” trials. This is an important distinction, since the initiation rate of neoadjuvant hormonal therapy is expected to be close to 100% because it is a prerequisite to the surgery. Third, the cohort consisted of the patients who were prescribed hormonal therapy medication; however, it would be useful to compare this population with patients who had the same disease but did not undergo hormonal therapy. Such a comparison could provide greater insight into a patient’s decision-making process with respect to this treatment. Fourth, while our study demonstrated the effectiveness of applying word clustering to group single terms, it is limited in that there were still several groups that were difficult to interpret their meanings. It will be interesting to investigate the extent to which including human judgment in clustering process can help improve the interpretability of word groups. Further, we identified the start of the hormonal therapy based on the medication entry date recorded in EMR system. It will be very interesting to investigate the intended start of the therapy. Future work can also consider incorporate insurance status, income and healthcare provider characteristics into the model, as well as investigate the influence that patient have on their first hormonal therapy prescription. Finally, another useful future direction is investigating the extent to which hormonal therapy initiation is associated with hormonal therapy medication discontinuation.

Acknowledgment

This work was supported by the National Science Foundation grant number IIS1418504.

References

Training a Convolutional Neural Network with Terminology Summarization Data Improves SNOMED CT Enrichment

Ling Zheng, PhD1, Hao Liu, MS2, Yehoshua Perl, PhD2, James Geller, PhD2
1CSSE Department, Monmouth University, West Long Branch, NJ, USA; 2SABOC Center, Department of Computer Science, NJIT, Newark, NJ, USA

Abstract
As a step toward learning to automatically insert new concepts into a large biomedical ontology, we are studying the easier problem of automatically verifying that an IS-A link should exist between a new child concept and an existing parent concept. We are using a Convolutional Neural Network, a powerful machine learning method. However, results depend on the quality of the training data. We use SNOMED CT (July 2017) for training and the subsequent release for testing. The main problem is to find a good set of negative training data. We experiment with two approaches, based on uncle-nephew (not connected) pairs of concepts. We contrast using the complete Clinical Finding hierarchy of SNOMED CT with using the powerful Area Taxonomy ontology summarization mechanism to constrain the training data. The results for the task of verifying IS-A links are improved by 8.6% when going from the complete hierarchy to the Area Taxonomy.

Introduction
Biomedical terminologies are structured and organized knowledge systems. They play a major role in enabling precise communication and in support of healthcare applications, e.g., Electronic Healthcare Record (EHR) systems [1-3]. A terminology, which typically consists of up to hundreds of thousands of concepts interconnected by many more relationships, can be viewed as a complex network with concepts as nodes and relationships as links. The hierarchical IS-A relationships between concepts are the backbone of the terminology. An IS-A relationship, from concept A to concept B, represents that concept A is a specialization of concept B. For example, within SNOMED CT [4], the concept Corneal infection IS-A Eye infection, because corneal infection is a specialization of eye infection. Due to the size and complexity of biomedical terminologies, their enrichment, i.e., adding new concepts at the proper positions in the concept hierarchy, is difficult and labor-intensive. Any semi-automatic framework that can predict/verify the place of a new concept in the terminology hierarchy is highly desirable.

In a previous study [5], we trained a Convolutional Neural Network (CNN) model to learn the likely locations of new IS-A links from the whole SNOMED CT hierarchy. We trained the model with data from the SNOMED CT 2017 July release and tested its ability to verify IS-A relationships of new concepts added to the SNOMED CT 2018 January release. This approach was based on document embeddings for each concept in the SNOMED CT. This preliminary study achieved an average F1 score of 0.70. In this paper, we suggest an improvement by using better training data.

One of the challenges when training machine learning (ML) models, and especially Deep Learning (DL) models, is dealing with imbalanced training data of different levels of quality. The training data is considered as high quality if it is balanced across the testing categories and is correlated to the test cases with good coverage. For example, to train a Convolutional Neural Network model for image classification, the data used for training has to be balanced across the image categories. It should be general enough to cover the possible test cases and task-oriented enough so that it does not deviate too far from the test data in various aspects, such as format, resolution, and brightness. High quality, task-oriented training data typically improves the results of ML techniques.

While training the CNN model with the whole SNOMED CT seemed to be a straightforward solution in our previous work, the modeling differences among the 19 hierarchies of SNOMED CT propagate inevitably to the CNN model. This may impair the performance of the trained CNN model in distinguishing IS-A links in different hierarchies. The features that the CNN model incorporates from the Clinical Finding hierarchy are rarely useful for distinguishing IS-A links in other hierarchies, e.g., Specimen, that are covering a different subject, because different hierarchies are modeled following different modeling principles and are often modeled by different curators.

Secondly, the training data provided by SNOMED CT is naturally unbalanced, as there are orders of magnitude fewer concept pairs connected by IS-A links than pairs of concepts not connected by IS-A links. This issue was handled by random downsampling from the set of concept pairs not connected by IS-A links. When selecting such pairs, we followed the idea of choosing negative training data that is closely related to the positive training data to
define an improved decision boundary. In particular, if B is a new concept with an IS-A relationship to a concept A, written as (B, A), then we select as the negative sample a pair of concepts B and C, such that C is a sibling of A and there is no IS-A relationship from B to C. This pair of concepts defines an uncle-nephew relationship relative to the pair (B, A).

In a long-range research program, our team has developed a taxonomy-based framework to support the summarization and visualization of biomedical ontologies [6-8]. An Area Taxonomy is a summarization network composed of area “nodes,” each representing a set of similar concepts. Training a model with uncle-nephew concept pairs from the same area implies that the CNN model will tend to incorporate more fine-grained features, learned from closer concept pairs, than general features when training with uncle-nephew pairs. This training data enables the CNN model to learn subtle differences between IS-A and non-IS-A pairs of concepts. Area taxonomy can also be utilized for quality assurance tasks, such as verifying missing IS-As, for a terminology [9].

Previous work on applying machine learning techniques to terminology enrichment is quite diverse. An effective weakly supervised ontology enrichment algorithm has been reported by Pembeci et al. [10] that uses concept similarity scores computed via Word2vec models to discover other related concepts. Jayawardana et al. used Word vector embeddings to derive candidate vectors and then trained an SVM model to calculate representative vectors for concepts [11]. OmicsCNN [12] is a dedicated DL framework to generate feature distances on “omics data” through a novel layer called OmicsConv to prepare omics data for clinical applications and scientific discovery.

In this work, we use the Clinical Finding hierarchy of 109,366 concepts of SNOMED CT as our test-bed. We utilize Doc2Vec [13] as our embedding encoder to generate concept-level vectors. We prepared two sets of training data, one using random SNOMED CT hierarchy data and another one with data obtained from our Area Taxonomy, and we trained two Convolutional Neural Network models respectively. We compared their performance on predicting the hierarchical IS-A relationships within the Clinical finding hierarchy of SNOMED CT. To validate our approach, we tested the two models with the 1,027 new concepts that were added to the Clinical finding hierarchy between the SNOMED CT July 2017 release and the January 2018 release.

Background

Convolutional Neural Network

Deep Learning [14] has been successfully applied to many tasks, such as computer vision [15] and speech recognition [16]. Convolutional Neural Networks (CNNs) have been extensively studied and used for various applications including image recognition, semantic parsing, search query retrieval, sentence modeling, classification, prediction, etc. CNNs utilize layers of convolving filters that are applied to local features [17], to serve as input for trainable classifiers for prediction tasks. This paper focuses on using a CNN as the model to solve a problem that is essential to ontology maintenance, namely how to verify the parent(s) for a newly added concept B. This is equivalent to verifying whether there is an IS-A link from a new concept B to an existing concept A.

Doc2vec

As most of the ML algorithms require input in numeric format, numeric representations of variable-length text, ranging from sentences to documents is a challenging task. Doc2vec (Paragraph Vectors) [13], an extension of Word2vec (word embedding) [18], projects variable-length documents or paragraphs to fixed-length vectors. It is an unsupervised framework that learns continuous distributed vector representations from unlabeled text data of a document, while preserving the inter-relationships within the text in the numeric format. In such vector representations, similar pieces of text are close to each other in Euclidean or cosine distance in lower dimensional vector spaces. The original Paragraph Vector paper [13] recommended combining the Distributed Memory version of Paragraph Vector (PV-DM) with the Distributed Bag of Words (PV-DBOW) representation to obtain consistency, which is done in this study.

Areas and Area Taxonomy

In previous work, we defined an area is a group of concepts all of which have identical sets of lateral relationship types. Lateral relationships are all binary relationships in SNOMED CT except for IS-A relationships. An Area Taxonomy is a network that consists of nodes representing areas and child-of links connecting areas. Area taxonomies have applications in quality assurance of terminologies [19]. Figure 1 illustrates the derivation of an Area Taxonomy. Figure 1(a) shows an excerpt of 16 concepts, drawn as labeled ovals, from SNOMED CT’s Clinical Finding hierarchy. Concepts of each area are grouped together in a dashed rectangle, labeled by the common set of relationships types. The arrows denote IS-A links. Lateral relationships are inherited along the IS-A
links between concepts. For example, *Infected corneal abrasion* and *Infective uveitis* (and three other concepts) are grouped together (in a green rectangle), because they all have the same lateral relationships: Associated morphology, Finding site, and Pathological process. Figure 1(b) shows the Area Taxonomy for 1(a). All colored rectangles represent nodes, which are connected by hierarchical child-of links that are derived from the corresponding IS-A relationships in SNOMED CT. Ochs et al. developed software [20] to automatically derive Area Taxonomies.

**Figure 1.** Derivation of Area Taxonomy. (a) Excerpt of 16 concepts from SNOMED CT’s Clinical Finding hierarchy. (b) Area Taxonomy for the excerpt in (a).

Areas with the same number of lateral relationship types have the same color and are shown at the same level. For example, the areas *{Causative agent, Finding site, Pathological process}* and *{Associated morphology, Finding site, Pathological process}* appear in the second level in green. *Eye infection* and its descendants in the grey rectangle are now represented by the area *{Finding site, Pathological process}*.

**Figure 2:** Serializing the hierarchical structure of one concept into one document.
Methods

We used the SNOMED CT July 2017 release as training set and the subsequent January 2018 release as testbed. The data preprocessing follows the same workflow as in our previous study [5] to generate documents and train embedding vectors, but is limited to the Clinical Finding hierarchy of SNOMED CT. We will illustrate the document generation process below, as it is crucial for understanding the next step.

i. Create one “document” per focus concept (see Figure 2). We use a concept’s corresponding SNOMED CT ID as the unique ID for a document. The hierarchical information related to this focus concept is used as the content of this document. Specifically, we chose this concept’s parents (targets of IS-A links from this concept), and its children (sources of IS-A links to this concept). Unlike our previous study, a concept’s sibling(s) are neglected in this study without showing accuracy loss. By constructing a document this way, a concept becomes the focus topic of this document and its closely related concepts “describe” the meaning of it in the ontology hierarchy.

ii. To maintain the existing hierarchical relationships of the focus concept to some degree, we arrange these concepts as follows: A concept’s parent(s) are placed at the beginning of the text, followed by the concept itself, and then by the child(ren) of this concept. Thus, the order of the whole document text is: Parent(s) — Focus concept — Child(ren) (Figure 2). Tabs are used as separators between concept groups. For example, Infectious thyroiditis has two parents, Thyroid infection and Thyroiditis, and two children, Acute suppurative thyroiditis and Viral thyroiditis. Thus, the content of the document for the concept Infectious thyroiditis is defined as “ID Thyroid infection, Thyroiditis \ Infectious thyroiditis \ Acute suppurative thyroiditis, Viral thyroiditis” where \ is the ASCII tab character, and ID stands for the ID of the focus concept.

Figure 3: Data flow during vectorization phase. N = 109,366 is the total number of concepts in the Clinical Finding hierarchy of the SNOMED CT July 2017 release.

After constructing a document for each concept, we passed the list of all documents to the Gensim implementation of Doc2Vec to generate two embeddings (displayed as black arrows in Figure 3). Gensim automatically generates
the document-level vectors while maintaining the semantic topics of the documents. We performed experiments with both the Distributed Memory version (PV-DM) and the Distributed Bag of Words (PV-DBOW) version of Paragraph Vector [13]. The PV-DM model and the PV-DBOW model function as two “vector dictionaries.” For each concept, we can retrieve its corresponding two vectors (denoted as Vector_{pv-dm} and Vector_{pv-dbow} in Figure 3) by sending its corresponding document ID to the two models, respectively. This is indicated by two blue arrows emanating from the green SNOMED CT repository. The vector dimension was set to 128. The use of those two embeddings is new and was not done in the previous study [5].

The same CNN model architecture was employed as in the previous study [5]. It was tailored to accommodate two vectors with a dimension of 128. The modification of stacking two vectors in depth for each concept as input does not require changing the previous CNN model, because the convolutional operations work across the depth dimension, such as, for example, when processing RGB channel data for color image input [14].

Distinguishing between IS-A and Non-IS-A Training Data

To train a CNN model that can verify an IS-A link between a pair of concepts, we need to train the model with both IS-A connected concept pairs and concept pairs with no IS-A connection. The IS-A connected concept pairs are explicitly defined in the terminology hierarchy. The judicious selection of non-IS-A concept pairs is critical for the accuracy of the model. In our previous study [5], we limited the non-IS-A pairs to only uncle-nephew pairs. Figure 4(a) illustrates this. Consider an IS-A relationship from a concept A to a concept B. A sibling C of B is an uncle of A. Thus the pair of A and C is called an uncle-nephew pair. The concept D has an IS-A relationship to concept Q and is not an uncle of A. The concept C is hierarchically related to B as its sibling. Thus, C has a higher probability of being mistaken as A's parent than D, which is not hierarchically related to B. In other words, the non-IS-A pair of A and C is more similar than the pair of unconnected concepts A and D to the IS-A pair (A, B). It is more useful to learn to distinguish between an IS-A pair and (one of) its uncle-nephew non-IS-A pairs than between an IS-A pair and an arbitrary non-IS-A pair, which might be “far away" from the IS-A pair.

![Figure 4](image-url) Illustrating two potential training data patterns

Utilizing the Area Taxonomy, the uncle-nephew pairs can be further divided into two types: uncle and nephew concepts are from the same area or uncle and nephew concepts are from different areas. In Figure 4(b), both C and D are A’s uncles, because they are siblings of B. However, C is in the same (red) area as A, while D resides in another (green) area. The uncle-nephew pair of A and C is more similar to the IS-A pair (A, B) than the uncle-nephew pair of A and D, since concepts from the same area share the same set of relationships types, and are more similar to each other. Our hypothesis is that the CNN model will become more accurate by training it with uncle-nephew pairs from the same area, because it can learn to distinguish between IS-A pairs and closely related non-IS-A pairs. As a result, the derived CNN model should be better at verifying whether a concept pair should be connected by an IS-A link or not, achieving better accuracy.

We demonstrate the above observation with a concrete example (Figure 5) from the Clinical Finding hierarchy. Let the nephew concept be Conjunctival diphtheria (in yellow). Its uncle concepts are Viral conjunctivitis, Parasitic conjunctivitis, and Tumorlet. Conjunctivitis (called also pinkeye) is an infection of the conjunctiva, the transparent tissue that covers the white part of the eye. These 3 concepts are the siblings of the concept Bacterial conjunctivitis, which is the parent of Conjunctival diphtheria. As Figure 5 shows, the uncle concept Viral conjunctivitis (in green) is in the same area {Associated morphology, Causative agent, Finding site, Pathological process} as Conjunctival diphtheria, while the other two uncle concepts Parasitic conjunctivitis and Tumorlet (in red) are in different areas.
Causative agent, Finding site, Pathological process] and [Associated morphology, Finding site, Pathological process], respectively, than Conjunctival diphtheria. Conjunctival diphtheria is structurally more similar to Viral conjunctivitis as both of them have the same four lateral relationships, while it is structurally different from Parasitic conjunctivitis and Tumorlet, which have only three lateral relationships.

Indeed, Viral conjunctivitis is closer to Bacterial conjunctivitis, the parent of Conjunctival diphtheria, than to Parasitic conjunctivitis and Tumorlet. Thus, we formulate the following hypothesis.

**Hypothesis:** Using uncle-nephew pairs as training data for non-IS-A pairs, with both taken from the same area, will improve the CNN model’s classification accuracy for non-IS-A and IS-A test cases, compared to training with pairs where uncle and nephew are randomly chosen from the Clinical Finding hierarchy.

**Testing the Hypothesis**

To evaluate our model’s performance on real, previously unseen data, we used new concepts from the SNOMED CT 2018 January release. A pre-trained Doc2Vec model returned a pre-determined vector for an existing concept or inferred a vector for a new concept from the new concept’s name. However, if all the word(s) in a new concept’s name are not in the trained vocabulary (i.e., they were never seen before when training the model), the inferred vector for this concept would be a random vector. To avoid this situation, we limited testing the CNN model to new concepts with multiple parents existing in the July 2017 release. It is true this limitation restricts our model’s applicability to about one-third of the concepts in SNOMED CT. For new concepts with only one existing parent, we suggest that it is better to use them for training the model afterward, rather than testing against them without having the model learned any knowledge about them. The detailed procedures for preparing test data for both IS-A concept pairs and non-IS-A concept pairs are as follows (Figure 6).

We used new concepts added in the SNOMED CT January 2018 release and their parents. To get the inferred vector for a new concept N, we randomly selected one of its parents – say Parent 1 in Figure 6(a). Then Parent 1’s name is concatenated with N’s name as the “query” to the Doc2Vec model. The vector inferred by the Doc2Vec model is then paired with one of the remaining parents’ vectors, say the vector for Parent 2 to form a vector pair (N, Parent 2). This vector pair, as one test case, is sent to the pre-trained CNN model for IS-A link verification. Similarly, we paired the inferred vector for N with vectors for Parent 3 and Parent 4, respectively to create two test cases, namely, whether the CNN model can verify the IS-A links for the pairs (N, Parent 3) and (N, Parent 4).

As the Doc2Vec model is pre-trained, i.e., it is not updated during the inference process, the inferred vectors are independent of each other. Therefore, we can create other sets of test cases by generating all the possible pairs between N and its parents. For example, as shown in Figure 6(b), we concatenated the names of N and Parent 2 to...
get the inferred vector for N, and then paired the inferred vector with vectors for Parent 1, Parent 3, and Parent 4 as three test cases.

Figure 6. (a) Three IS-A test cases if pairing up the new concept N with its first parent. (b) Repetition of (a), but now we are pairing up N with Parent 2.

Following this setup, out of 2,005 new concepts added to the Clinical Finding hierarchy of the January 2018 release, we selected 1,027 concepts to generate a total of 7,494 IS-A concept pairs for testing the CNN model’s accuracy when verifying IS-A links for the new SNOMED CT release. Out of the 1027 concepts, 797 concepts are leaf nodes.

For non-IS-A test pairs, we first collected the existing parents of all the new concepts that were used for IS-A testing to form a “Parent set.” Then for each new concept, we paired it with all the concepts from the “Parent set” except its own parents, generating non-IS-A test pairs (Figure 7). By doing this, the non-IS-A test data is not limited to only uncle-nephew pairs, but allows general combinations of new concepts with other existing parent concepts. The CNN model’s performance on such concept pairs reveals its generalizability to the broader population of non-IS-A test data. We randomly selected 7,494 non-IS-A concept pairs to match the number of IS-A concept pairs.

The process to convert from a test pair into a vector pair is demonstrated in Figure 7. After the test pairs for both positive and negative samples were selected, we query the PV-DM and PV-DBOW models to get the corresponding concept vectors. For each concept pair, we obtained four vectors (two vectors for each concept). To get the corresponding vector pair, we concatenated the two vectors generated by the same model into one vector. Then the two long vectors were sent into the pre-trained CNN model in a random order to compute a class label (0 or 1) as the prediction result. For an IS-A test concept pair, label 1 is correct if there is an IS-A link between these two concepts in the new SNOMED CT release. For a non-IS-A concept pair, label 0 is correct if there is no IS-A link between these two concepts in the new SNOMED CT release. We repeated the training and testing ten times with randomly chosen 90:10 splits between training and validation data to obtain a consistent evaluation of the efficacy of our methodology.
Besides examining the quality of the training data, we also scrutinized the quality of the concept vectors from the Doc2Vec embedding. We experimented with the “iterations/epochs” parameters of the PV-DM and PV-DBOW models. This is a parameter to control how many times to iterate over the training corpus. The typical iteration counts suggested in the original “Paragraph Vectors” paper [13] are 10-20 for tens-of-thousands to millions of documents. Thus, we experimented with 10, 20, and 50 iterations for comparison.

![Data flow during testing CNN model with new data (use case)](image)

**Figure 7:** Data flow during testing CNN model with new data (use case)

**Results**

We first report the prediction results of our CNN model with the 7,494 IS-A connected concept pairs. Table 1 summarizes the performance comparison between using original (= “unsummarized”) training data from the whole Clinical Finding hierarchy (shown as “hierarchy”) versus training data from the same area only (shown as “area”). The Precision, Recall, and F1 scores for ten tests are presented in Table 1. The model trained using the Area Taxonomy data is superior to the model trained using hierarchy data in both best F1 score (0.78 vs. 0.73) and average F1 score (0.75 vs. 0.69). The standard deviations of the ten F1 scores for both trained models are the same (0.018).

**Table 1. IS-A testing results for ten tests, comparing SNOMED CT hierarchy training data with “in area” training data**

<table>
<thead>
<tr>
<th>Index</th>
<th>Precision hierarchy</th>
<th>Recall hierarchy</th>
<th>F1 hierarchy</th>
<th>Precision area</th>
<th>Recall area</th>
<th>F1 area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.85</td>
<td>0.59</td>
<td>0.65</td>
<td>0.69</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.86</td>
<td>0.63</td>
<td>0.70</td>
<td>0.73</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.86</td>
<td>0.58</td>
<td>0.67</td>
<td>0.69</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.87</td>
<td>0.59</td>
<td>0.74</td>
<td>0.70</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.86</td>
<td>0.58</td>
<td>0.64</td>
<td>0.69</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.86</td>
<td>0.58</td>
<td>0.69</td>
<td>0.69</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.87</td>
<td>0.57</td>
<td>0.71</td>
<td>0.69</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.87</td>
<td>0.53</td>
<td>0.76</td>
<td>0.66</td>
<td><strong>0.78</strong></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.85</td>
<td>0.62</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.86</td>
<td>0.58</td>
<td>0.69</td>
<td>0.69</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.86</td>
<td>0.59</td>
<td>0.69</td>
<td>0.69</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.007</td>
<td>0.022</td>
<td>0.041</td>
<td>0.018</td>
<td>0.018</td>
<td></td>
</tr>
</tbody>
</table>
Next, we evaluated our model’s performance on non-IS-A concept pairs. The Precision, Recall, and F1 scores for ten tests are presented in Table 2. Table 2 summarizes the performance comparison for non-IS-A concept pairs between using data from the whole hierarchy (shown as “hierarchy”) versus using training data from the same area (shown as “area”). The model trained with Area Taxonomy data achieved the same F1 value as the model trained with the whole hierarchy (0.78). The standard deviation of the ten F1 scores for the “hierarchy” model is 0.005, and it is 0.008 for the “area” model.

Table 2. Non-IS-A testing results among ten tests, comparing SNOMED CT hierarchy training data with “in area” training data

<table>
<thead>
<tr>
<th>Index</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hierarchy</td>
<td>area</td>
<td>hierarchy</td>
</tr>
<tr>
<td>1</td>
<td>0.69</td>
<td>0.71</td>
<td>0.89</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.74</td>
<td>0.90</td>
</tr>
<tr>
<td>3</td>
<td>0.68</td>
<td>0.72</td>
<td>0.90</td>
</tr>
<tr>
<td>4</td>
<td>0.69</td>
<td>0.76</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>0.68</td>
<td>0.71</td>
<td>0.90</td>
</tr>
<tr>
<td>6</td>
<td>0.68</td>
<td>0.73</td>
<td>0.90</td>
</tr>
<tr>
<td>7</td>
<td>0.68</td>
<td>0.74</td>
<td>0.91</td>
</tr>
<tr>
<td>8</td>
<td>0.66</td>
<td>0.77</td>
<td>0.92</td>
</tr>
<tr>
<td>9</td>
<td>0.70</td>
<td>0.71</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>0.68</td>
<td>0.73</td>
<td>0.90</td>
</tr>
<tr>
<td>Average</td>
<td>0.69</td>
<td>0.73</td>
<td>0.90</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.014</td>
<td>0.021</td>
<td>0.009</td>
</tr>
</tbody>
</table>

The training time for the 10, 20, and 50 iterations was 485, 947, and 2035 seconds, respectively, with the same hardware configuration. The CNN models trained on top of these three vector spaces achieved an average F1 score of approximately 0.745 for IS-A testing, for each of the 10, 20, and 50 iterations. For non-IS-A testing, the average F1 scores for the 10, 20, and 50 iterations were 0.784, 0.775, and 0.753, respectively.

Discussion

In this study, we set out to evaluate the hypothesis whether using terminology summarization, in the form of the Area Taxonomy summarization network, can help in obtaining high-quality training data. For this, we used only uncle-nephew pairs from the same area as the non-IS-A training data. The hypothesis was proven only partially correct. Sharpening the distinction between IS-A pairs and non-IS-A uncle-nephew pairs by selecting non-IS-A pairs from the same area improved the F1 score for identifying IS-A pairs by 8.6% on average from 0.69 to 0.75, while for non-IS-A pairs there was no improvement. Perhaps focusing the non-IS-A training on uncle-nephew pairs close to IS-A pairs, helps improve identifying IS-A pairs, but not the broad non-IS-A pair population tested. We note that the non-IS-A testing was set up by pairing each focus concept with the parents of all other focus concepts. That is, those are not uncle-nephew pairs. Such broad testing data is a fair representation of the large number of non-IS-A pairs. However, the non-IS-A training performed in our study, by its nature, did not improve identifying these non-IS-A pairs.

The average F1 score for non-IS-A pairs is 0.78, which is slightly higher than for IS-A pairs (0.75). The standard deviation is lower for the non-IS-A pairs. Interestingly, for IS-A pairs the recall is higher than the precision while it is the opposite for non-IS-A pairs. Even though the average F1 score (0.78) for non-IS-A did not improve with summarization, it is still higher than the F1 score (0.75) for IS-A identification. The corresponding best F1 scores are even closer (0.80 vs. 0.78). For the hierarchy approach the IS-A identification is weaker (average F1 =0.69) than identifying non-IS-A (average F1=0.78). Hence the area summarization approach improved the “weakest point” (identifying IS-A pairs) of the overall performance of the hierarchy approach by obtaining an average F1 of 0.75. As a result, the average F1 scores of both the IS-A and the non-IS-A testing are close (0.75 vs. 0.78) for the area summarization approach. Experimenting with 10, 20, and 50 iterations, the results indicated that while more iterations took more time they actually diminished the F1 values. Hence, there is no point of running more than 10 iterations since the test performance is not improving.
Conclusion

The complexity of terminological data poses a challenge to training machine learning models that can achieve high accuracy on classifying hierarchical relationships between new concepts and their parents. To overcome this challenge, we explored the use of data obtained from the Area Taxonomy, a terminology summarization network, to prepare high-quality training data that can better distinguish between IS-A and non-IS-A concept pairs. With SNOMED CT’s Clinical Finding hierarchy as a testbed, the CNN model trained with the summarized data achieved an average F1 score of 0.75 for identifying IS-A pairs, which is about 8.6% better than for the model trained without summarization.

Acknowledgements

Research reported in this publication was partially supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Leveraging Non-lattice Subgraphs to Audit Hierarchical Relations in NCI Thesaurus

Rashmie Abeysinghe\textsuperscript{1,2}, Michael A. Brooks, MD\textsuperscript{3}, Licong Cui, PhD\textsuperscript{1,*}

\textsuperscript{1}School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

\textsuperscript{2}Department of Computer Science, University of Kentucky, Lexington, KY

\textsuperscript{3}Departments of Radiology and Medicine, University of Kentucky, Lexington, KY

Abstract

Auditing National Cancer Institute (NCI) thesaurus is essential to ensure that it provides accurate terminology for cancer-related clinical care as well as translational and basic research. We leverage a structural-lexical approach to identify missing hierarchical IS-A relations in NCI thesaurus based on non-lattice subgraphs and derived lexical attributes of concepts. For each concept in a non-lattice subgraph, we use two ways to derive the concept’s lexical attributes: (1) inheriting lexical attributes from its ancestors within the subgraph; and (2) inheriting lexical attributes from all its ancestors. For a pair of concepts not having a hierarchical relation, if the lexical attributes of one concept is a subset of that of the other, we suggest there is a potential missing IS-A relation between the two concepts. Our approach identified 547 non-lattice subgraphs in the 19.01d release of NCI thesaurus which revealed a total of 1,022 unique potential missing IS-A relations. A random sample of 100 relations was evaluated by a domain expert. Among these relations, 90 can be obtained by the way of inheriting lexical attributes from ancestors within non-lattice subgraph, among which 76 were confirmed as valid (a precision of 84.44%); and 82 can be obtained by the way of inheriting all ancestors, among which 73 were confirmed as valid (a precision of 89.02%). The results show that our structural-lexical approach based on non-lattice subgraphs is effective for auditing NCI thesaurus.

1 Introduction

The National Cancer Institute thesaurus (NCIt) is a biomedical terminology produced by NCI Enterprise Vocabulary Services containing more than 140,000 concepts\textsuperscript{1}. Auditing biomedical terminologies such as NCIt is essential to ensure that it produces an accurate representation of the knowledge of the domain it models. This is especially important because the quality issues in terminologies would cause the applications that use these terminologies to be erroneous as well\textsuperscript{2}. As the terminologies are continuously being expanded, their complexity also increases, making the introduction of errors almost unavoidable. Therefore, Terminology Quality Assurance (TQA) has become an important part of the terminological management lifecycle of all modern biomedical terminologies. However, manually reviewing a terminology to perform TQA is impractical due to the increasing size and complexity of modern terminologies. Therefore, automated or semiautomated approaches are needed to perform TQA efficiently and effectively.

In this paper, we introduce a structural-lexical approach based on non-lattice subgraph (NLS) to identify missing hierarchical relations in NCIt. Recently a number of studies have shown that analyzing lexical features in concept labels in NLSs is a promising way to identify different kinds of defects in biomedical terminologies\textsuperscript{3–5}. In this work, we further work on this idea to come up with a set of lexical attributes for each concept in an NLS. We leverage these lexical attributes to suggest potential missing hierarchical relations in the NLS. A domain expert reviewed a randomly selected sample from the potential missing relations derived to examine the effectiveness of our approach.

2 Background

2.1 NCI Thesaurus (NCIt)

NCI Thesaurus is a biomedical terminology which covers vocabulary for cancer-related clinical care, translational and basic research, and public information and administrative activities\textsuperscript{1}. It was originally created to facilitate interoperability and data sharing by various components of NCI by incorporating terms used by different components and mapping them to unique concepts\textsuperscript{6}. Each concept includes a unique code, a preferred term, abbreviations, synonyms, and definitions\textsuperscript{7}. The content of NCIt is organized in a description logic environment with more than 400,000 relations

\*Corresponding author. Email: licong.cui@uth.tmc.edu
between concepts. NCIt is updated monthly with around 700 new concepts and many additional changes in each new release. It is released in many formats including Ontylog XML, OWL, and flat files. It is also available in both defined and inferred versions. We used the inferred version of the 19.01d release of NCIt in OWL format in this work.

2.2 Terminology Quality Assurance

Various methods have been investigated to facilitate quality improvement of biomedical terminologies like NCIt\(^8\). A type of high-level summary graphs, called abstraction networks, have been widely used to find inconsistencies in many biomedical terminologies\(^9-12\). Min et al.\(^13\) have applied such an approach to the Biological Process hierarchy of NCIt which has led to the identification of different types of errors such as missing roles, missing concepts, incorrect hierarchical relations etc. Mougin et al.\(^14\) have utilized the relations in UMLS semantic network to audit hierarchical and associative relations in NCIt. He et al.\(^15,16\) have employed topological patterns that exists between NCIt and a reference terminology to import new concepts to NCIt. In previous work\(^17,18\), we introduced a lexical-based inference approach to detect missing and incorrect relations in the Gene Ontology. Zheng et al.\(^19\) have proposed a deep learning-based approach to predict concept names for new concepts that are added to SNOMED CT. Cui et al. investigated different ways of analyzing lexical features of concepts in NLSs to uncover missing hierarchical relations in SNOMED CT\(^3,5\) and in NCIt\(^4\).

2.3 Non-Lattice Subgraphs (NLSs)

Being a lattice is considered a desirable property for a well-formed terminology\(^20\). A terminology is a lattice if any pair of concepts has a unique maximal common descendant and a unique minimal common ancestor. Here, a common descendant \(C\) is known as a maximal common descendant of a concept-pair \((A, B)\), if \(A\) and \(B\) have no other common descendant \(D\) such that \(C\) is a descendant of \(D\); and similarly, a common ancestor \(R\) is known as a minimal common ancestor of a concept-pair \((P, Q)\), if \(P\) and \(Q\) have no other common ancestor \(S\) such that \(R\) is an ancestor of \(S\). If a concept pair has more than a single maximal common descendant (or a single minimal common ancestor), it is known as a non-lattice pair\(^20,21\), which may reveal quality issues in terminologies.

It is not economical to separately examine multiple non-lattice pairs which share the same maximal common descendants. To address this, non-lattice subgraphs (NLSs) have been introduced\(^3\). An NLS can be acquired by a non-lattice pair \((c_1, c_2)\) as follows. Firstly, maximal common descendants of the non-lattice pair, \(mcd(c_1, c_2)\), named as the lower bounds, is computed. Then, the minimal common ancestors of the lower bounds, \(mca(mcd(c_1, c_2))\), named as the upper bounds, is computed. Finally, all the concepts as well as relations between (and including) lower and upper bounds is aggregated to generate the NLS. The size of an NLS is the number of concepts it contains. For example, in Figure 1, the non-lattice pair \{1, 2\} (alternatively \{1, 3\} or \{2, 3\}) yields \{6, 7\} as its maximal common descendants. Reversely computing minimal common ancestors of \{6, 7\} yields \{1, 2, 3\}. Then, the concepts \{4, 5\} as well as relations between \{1, 2, 3\} and \{6, 7\} are aggregated to form the given NLS.

\[
mca(mcd(1,2))
\]

\[
mcd(1, 2)
\]

\[
mcd(1, 2)
\]

**Figure 1:** An example of an NLS. Nodes of the graph are concepts. The edges indicate hierarchical IS-A relations where the arrowheads point to the parent concept.

NLSs have been utilized to effectively identify defects in biomedical terminologies. Cui et al.\(^3\) have proposed four lexical patterns found in NLSs which suggest missing hierarchical relations and missing concepts in SNOMED CT. In a previous work\(^4\), we introduced two new lexical patterns applying that approach to NCIt. Cui et al.\(^5\) originally introduced an approach combining NLSs and enriched lexical attributes of concepts to identify missing and incorrect
hierarchical relations in SNOMED CT. We also introduced a method to identify similar NLSs in the Gene Ontology to reduce the effort needed by domain experts in reviewing them.

3 Methods

We first extract all the NLSs in the 19.01d release of NCIt. Then we construct the lexical attributes of concepts in NLSs by two ways: (1) inheriting lexical attributes from ancestors within NLSs; and (2) inheriting lexical attributes from all the ancestors. Based on the lexical attributes, we identify potential missing hierarchical relations between concepts. A random sample of missing relations is evaluated by a domain expert to verify their correctness.

3.1 Constructing Lexical attributes of Concepts

Two lexical sources are leveraged to construct the set of lexical attributes for each concept in an NLS. Firstly we consider lexical attributes of the concept itself. The second source relies on the lexical attributes of the concept’s ancestors. The second source is obtained in two ways.

1. Inheriting lexical attributes from ancestors within the NLS: In this way, we consider concept’s ancestors that reside within the NLS to enrich the lexical attributes of a particular concept in the NLS. Note that we consider all the direct and indirect ancestors of a concept. Therefore, we compute the transitive closure of the hierarchical relation within the NLS to obtain indirect (transitive) ancestors.

2. Inheriting lexical attributes from all the ancestors: In this way, we consider all the concept’s ancestors in the terminology without limiting to the NLS. To obtain indirect (transitive) ancestors, we compute the transitive closure of the hierarchical relation in the entire terminology.

We compare these two ways later in the paper in Section 4. Using these two sources we construct a set of lexical attributes $L_c$ for each concept $c$ in an NLS as follows.

- Load $L_c$ with the set of words contained in the preferred name of $c$.
- For each ancestor $a$ of $c$, add the set of words contained in the preferred name of $a$ to $L_c$. Note that $a$ could be an ancestor within the NLS or an ancestor external to the NLS depending on which way is used as discussed above.

![Diagram]

**Figure 2:** An NLS of size 4 and its remediation. The suggested remediation here is a missing hierarchical relation: “C21663: Fibroadenoma of the Mouse Mammary Gland” IS-A “C21665: Adenoma of the Mouse Mammary Gland”. This can be obtained by both ways: inheriting lexical attributes from ancestors within the NLS and from all the ancestors.

We demonstrate the construction process using the NLS shown in Figure 2, considering ancestors within the NLS. For each concept $c$ in the NLS, we construct a set of attributes $L_{w_c}$ as follows. We initialize $L_{w_c}$ with the lexical attributes obtained from $c$’s preferred name:

- $L_{w_1} = \{\text{mouse, adenoma}\}$
- $L_{w_2} = \{\text{benign, neoplasms, of, the, mouse, mammary, gland}\}$
- $L_{w_3} = \{\text{adenoma, of, the, mouse, mammary, gland}\}$
- $L_{w_4} = \{\text{fibroadenoma, of, the, mouse, mammary, gland}\}$
If we enrich the above sets with the lexical attributes of the ancestors within the NLS, then the resulting attribute sets for each concept \( c (L_{w_i}) \) are as follows (newly added attributes are underlined):

\[
\begin{align*}
L_{w_1} &= \{ \text{mouse, adenoma} \} \\
L_{w_2} &= \{ \text{benign, neoplasms, of, the, mouse, mammary, gland} \} \\
L_{w_3} &= \{ \text{adenoma, of, the, mouse, mammary, gland, benign, neoplasms} \} \\
L_{w_4} &= \{ \text{fibroadenoma, of, the, mouse, mammary, gland, adenoma, benign, neoplasms} \}
\end{align*}
\]

If we use the lexical attributes of all the ancestors in the terminology, then the resulting attribute sets for each concept \( c (L_{a_i}) \) are as follows:

\[
\begin{align*}
L_{a_1} &= \{ \text{mouse, adenoma, murine, organism, benign, epithelial, diagnosis, neoplasm, experimental, neoplasms, cell} \} \\
L_{a_2} &= \{ \text{benign, neoplasms, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, experimental, murine, disorder, system, neoplasm} \} \\
L_{a_3} &= \{ \text{adenoma, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, epithelial, experimental, cell, murine, disorder, system, benign, neoplasm, neoplasms} \} \\
L_{a_4} &= \{ \text{fibroadenoma, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, epithelial, experimental, cell, murine, disorder, system, adenoma, benign, neoplasm, neoplasms} \}
\end{align*}
\]

### 3.2 Detecting Missing Relations

For a pair of concepts \( c_1 \) and \( c_2 \) in an NLS that are not connected by a hierarchical relation, if \( c_2 \)'s lexical attributes \( L_{c_2} \) is a proper subset of the \( c_1 \)'s lexical attributes \( L_{c_1} \), then we suggest \( c_1 \) IS-A \( c_2 \) (i.e. \( c_1 \) is the more specific concept).

After obtaining all such potential missing relations in an NLS, we remove redundant relations that can be inferred by others. For example, if we suggest \( a \) IS-A \( b \) and \( a \) IS-A \( c \) for a particular NLS where \( b \) IS-A \( c \) already exists in the NLS, then we consider \( a \) IS-A \( c \) as redundant, since it can be inferred transitively through \( a \) IS-A \( b \) and \( b \) IS-A \( c \). Therefore, we remove \( a \) IS-A \( c \) from the list of suggestions.

For instance, considering ancestors within NLS, for concepts 3 and 4 in Figure 2, \( L_{w_1} = \{ \text{adenoma, of, the, mouse, mammary, gland, benign, neoplasms} \} \) is a proper subset of \( L_{w_4} = \{ \text{fibroadenoma, of, the, mouse, mammary, gland, adenoma, benign, neoplasms} \} \). Also, considering all the ancestors, \( L_{a_3} = \{ \text{adenoma, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, epithelial, experimental, cell, murine, disorder, system, benign, neoplasm, neoplasms} \} \) is a proper subset of \( L_{a_4} = \{ \text{fibroadenoma, of, the, mouse, mammary, gland, integumentary, organism, diagnosis, epithelial, experimental, cell, murine, disorder, system, adenoma, benign, neoplasm, neoplasms} \} \).

Hence, we suggest concept 4 should be more specific than 3, i.e. Fibroadenoma of the Mouse Mammary Gland IS-A Adenoma of the Mouse Mammary Gland. As discussed above, this can be obtained by both considering ancestors within the NLS and all the ancestors.

Figure 3 contains a size-7 NLS with a potential missing hierarchical relation: “C4887: Metastatic Malignant Neoplasm in the Trachea” IS-A “C4571: Malignant Respiratory Tract Neoplasm” which can be obtained only by considering ancestors within the NLS for constructing lexical attributes.

Figure 4 contains a size-11 NLS with a potential missing hierarchical relation: “C5270: Cerebellar Papillary Meningioma” IS-A “C3569: Malignant Cerebellar Neoplasm” which can be only obtained by considering all the ancestors for constructing lexical attributes. This is because inheriting lexical attributes from ancestors within the NLS yields \( L_{w_11} = \{ \text{malignant, cerebellar, neoplasm, infratentorial, brain, intracranial, central, nervous, system} \} \) which is not a proper subset of \( L_{w_10} = \{ \text{cerebellar, papillary, meningioma, grade, iii, malignant, neoplasm} \} \). However, when all the ancestors are considered, \( L_{a_{11}} = \{ \text{malignant, cerebellar, neoplasm, disorder, central, system, nervous, infratentorial, intracranial, brain} \} \) is a subset of \( L_{a_{10}} = \{ \text{cerebellar, papillary, meningioma, infratentorial, intracranial, brain, cell, malignant, disorder, system, central, meningeal, nervous, grade, iii, neoplasm, meningothe} \} \).
Figure 3: An NLS of size 6 and its remediation. The suggested remediation here is a missing hierarchical relation: “C4887: Metastatic Malignant Neoplasm in the Trachea” IS-A “C4571: Malignant Respiratory Tract Neoplasm”. This can only be obtained by considering the ancestors within the NLS for enriching lexical attributes.

Figure 4: An NLS of size 11 and its remediation. The suggested remediation here is a missing hierarchical relation: “C5270: Cerebellar Papillary Meningioma” IS-A “C3569: Malignant Cerebellar Neoplasm”. This can only be obtained by considering all ancestors for enriching lexical attributes.

3.3 Filtering

We perform three kinds of filtering to avoid generate erroneous suggestions of potential missing IS-A relations: stop word filtering, antonym filtering, and position filtering.

Stop word filtering. Consider the concepts “C4013: Malignant Head and Neck Neoplasm” and “C3260: Neck Neoplasm”. These two satisfy all the requirements to be candidates for a suggestion of a missing hierarchical relations in the form of “C4013: Malignant Head and Neck Neoplasm” IS-A “C3260: Neck Neoplasm”. However, upon close
observation, it can be seen that this suggestion is wrong since it gives the idea of *Head Neoplasm* is a subtype of *Neck Neoplasm*. Existence of such stop words in concepts make them more prone to generate erroneous missing hierarchical relation suggestions. Therefore, If a concept contains such stop words, we do not make any suggestions. Moreover, we also do not consider such concepts to enrich lexical attributes of other concepts. That is, if a concept with stop words exists as an ancestor of another concept, we do not enrich the lexical attributes of the latter with the former. The stop words used to perform this filtering are: “and”, “and/or”, “or”, “no”, “not”, “without”, “due to”, “secondary to”, “except”, “by”, “after”, “able”, “removal”, “replacement”, “NOS”, where “NOS” represents “Not Otherwise Specified”.

**Antonym filtering.** If the constructed enriched lexical attributes of a particular concept contains an antonym pair, such concepts are more prone to erroneous suggestions as well. For example, consider the concepts “C60996: Malignant Epithelial Small Polygonal Cell” with attributes {small, cytoplasm, with, large, abundant, polygonal, epithelial, neoplastic, cell, malignant} and “C36822: Malignant Epithelial Large Cell” with attributes {large, epithelial, neoplastic, cell, malignant}. Even though attributes of C60996 is a proper subset of C36822, suggesting a hierarchical relation between these two is obviously not accurate since C60996 is discussing small cells and C36822 is discussing large cells (note that “small” and “large” is an antonym pair). Therefore, after obtaining the set of attributes, we check the set to ensure that it does not contain an antonym pair. The antonym pairs are obtained from WordNet.

**Position filtering.** For concepts with short names, they may appear as a part of other concepts’ names in various positions (e.g., beginning, middle, or end). For concepts whose names are not appearing at the end of other concepts’ names, it is likely to suggest incorrect missing IS-A relations. For instance, concept “Fentanyl” appears at the beginning of concept “Fentanyl Citrate Pectin-Based Nasal Spray”, and the subset inclusion may wrongly suggest “Fentanyl Citrate Pectin-Based Nasal Spray” IS-A “C494:Fentanyl”. Therefore, we filter out such cases by assigning a constraint such that the shorter concept should always appear at the end of the longer concept.

### 3.4 Evaluation

To evaluate the performance of our approach in accurately identifying missing hierarchical relations, we randomly selected a sample of missing hierarchical relations from the overall results for evaluation. These samples were provided to a domain expert (author MAB). Existing erroneous hierarchical relations in NCIt may help derive incorrect suggestions for missing hierarchical relations. Therefore, for the potential missing relations identified as incorrect by the domain expert, in a second round of evaluation, we provided the domain expert with existing hierarchical relations that were used to derive the incorrect ones. If the domain expert disagrees with the existing relation as well, then we marked it as an incorrect existing hierarchical relation. For instance, the NLS in Figure 5 denotes such a scenario.

**Figure 5:** An NLS of size 8 and its remediation. The suggested remediation here is an incorrect hierarchical relation: “C3779: Giant Cell Carcinoma” should not be a subtype of “C3780: Large Cell Carcinoma”. The existence of this relation derives the incorrect suggestion of “C4452: Lung Giant Cell Carcinoma” IS-A “C4450: Lung Large Cell Carcinoma”.

“Ancient” filtering. If a concept contains such stop words, we do not make any suggestions. Moreover, we also do not consider such concepts to enrich lexical attributes of other concepts. That is, if a concept with stop words exists as an ancestor of another concept, we do not enrich the lexical attributes of the latter with the former. The stop words used to perform this filtering are: “and”, “and/or”, “or”, “no”, “not”, “without”, “due to”, “secondary to”, “except”, “by”, “after”, “able”, “removal”, “replacement”, “NOS”, where “NOS” represents “Not Otherwise Specified”.

**Antonym filtering.** If the constructed enriched lexical attributes of a particular concept contains an antonym pair, such concepts are more prone to erroneous suggestions as well. For example, consider the concepts “C60996: Malignant Epithelial Small Polygonal Cell” with attributes {small, cytoplasm, with, large, abundant, polygonal, epithelial, neoplastic, cell, malignant} and “C36822: Malignant Epithelial Large Cell” with attributes {large, epithelial, neoplastic, cell, malignant}. Even though attributes of C60996 is a proper subset of C36822, suggesting a hierarchical relation between these two is obviously not accurate since C60996 is discussing small cells and C36822 is discussing large cells (note that “small” and “large” is an antonym pair). Therefore, after obtaining the set of attributes, we check the set to ensure that it does not contain an antonym pair. The antonym pairs are obtained from WordNet.

**Position filtering.** For concepts with short names, they may appear as a part of other concepts’ names in various positions (e.g., beginning, middle, or end). For concepts whose names are not appearing at the end of other concepts’ names, it is likely to suggest incorrect missing IS-A relations. For instance, concept “Fentanyl” appears at the beginning of concept “Fentanyl Citrate Pectin-Based Nasal Spray”, and the subset inclusion may wrongly suggest “Fentanyl Citrate Pectin-Based Nasal Spray” IS-A “C494:Fentanyl”. Therefore, we filter out such cases by assigning a constraint such that the shorter concept should always appear at the end of the longer concept.

### 3.4 Evaluation

To evaluate the performance of our approach in accurately identifying missing hierarchical relations, we randomly selected a sample of missing hierarchical relations from the overall results for evaluation. These samples were provided to a domain expert (author MAB). Existing erroneous hierarchical relations in NCIt may help derive incorrect suggestions for missing hierarchical relations. Therefore, for the potential missing relations identified as incorrect by the domain expert, in a second round of evaluation, we provided the domain expert with existing hierarchical relations that were used to derive the incorrect ones. If the domain expert disagrees with the existing relation as well, then we marked it as an incorrect existing hierarchical relation. For instance, the NLS in Figure 5 denotes such a scenario. “C3779: Giant Cell Carcinoma” should not be a subtype of “C3780: Large Cell Carcinoma”. The existence of this relation derives the incorrect suggestion of “C4452: Lung Giant Cell Carcinoma” IS-A “C4450: Lung Large Cell Carcinoma”.

![Figure 5](image-url)
4 Results

4.1 Summary Results

A total of 9,512 NLSs were extracted from the 19.01d version of NCIt with sizes ranging from 4 to 644. Out of these, our approach identified 547 NLSs with potential missing hierarchical relations. These NLSs contained a total of 1,022 potential missing hierarchical relations (note that an NLS may contain more than one missing hierarchical relation). It can be seen from Table 1 that 441 out of 547 NLSs can be identified by the way of inheriting lexical attributes from ancestors within NLS and suggests 925 potential missing IS-A relations; and 422 out of 547 can be identified by the way of inheriting lexical attributes from all the ancestors and suggests 847 potential missing IS-A relations. The two ways identified 750 potential missing IS-A relations in common.

Table 1: The number of NLSs and the number of potential missing hierarchical relations suggested in those NLSs.

<table>
<thead>
<tr>
<th>Type</th>
<th># of NLSs</th>
<th># of potential missing IS-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inheriting lexical attributes from ancestors within NLS</td>
<td>441</td>
<td>925</td>
</tr>
<tr>
<td>Inheriting lexical attributes from all the ancestors</td>
<td>422</td>
<td>847</td>
</tr>
</tbody>
</table>

4.2 Evaluation

The evaluation sample contained 100 potential missing relations observed in 83 NLSs identified by our approach. The domain expert concluded 85 (85%) of missing hierarchical relations are valid. Table 2 shows 15 examples of valid missing hierarchical relations in the form of subconcept and superconcept, for instance, “C7155: Primary Central Chondrosarcoma” IS-A “C3737: Mesenchymal Chondrosarcoma”. For the 15 invalid ones, the domain expert further inspected the existing hierarchical relations that were used to derive the invalid ones and verified that 8 of them were actually incorrect. Table 3 lists five examples of incorrect existing relations. For example, “C66775: Borderline Ovarian Mucinous Adenofibroma” should not be a subtype of “C4934: Benign Female Reproductive System Neoplasm”, since the word “borderline” indicates that it is on the borderline between benign and malignant, and may exhibit malignant behavior.

We summarize the evaluation result in Table 4 according to the two ways of inheriting lexical attributes. Among 100 NLSs, 90 were identified by the way of inheriting lexical attributes from ancestors within NLS suggesting 76 correct missing IS-A relations (a precision of 84.44%); and 82 were identified by the way of inheriting lexical attributes from all the ancestors suggesting 73 correct missing IS-A relations (a precision of 89.02%).

5 Discussion

This paper presents a structural-lexical approach to audit NCIt based on enriched lexical attributes of concepts in NLSs. The results indicate that most missing IS-A relations can be commonly obtained by considering ancestors within the NLSs and all the ancestors to enrich the lexical attributes. The former way identified more potential missing IS-A relations than the latter did, while the latter achieved a better precision than the former did.

5.1 Analysis of Failure Cases

The primary focus of this work was to identify missing hierarchical relations in NCIt. Upon observation of the false positives, it could be noted that a majority of them (53%) occur due to the existing erroneous hierarchical relations in NCIt. For example, in Figure 5, Giant Cell Carcinoma is categorized as a subtype of Large Cell Carcinoma. However, under the current (2015) WHO classification, Giant Cell Carcinomas are classified as a separate category of tumor. Therefore, Giant Cell Carcinoma should not be a subtype of Large Cell Carcinoma. Likewise in a separate case, our approach inaccurately identified “C39951: Testicular Fibroma” IS-A “C4092: Benign Epithelial Neoplasm” as a missing relation. However, it could be seen that this was obtained due to the erroneous existing relation “C39951: Testicular Fibroma” IS-A “C3709: Epithelial Neoplasm”, since a Testicular Fibroma does not arise from Testicular
Table 2: Fifteen examples of valid missing hierarchical relations obtained by our approach.

<table>
<thead>
<tr>
<th>Subconcept</th>
<th>Superconcept</th>
</tr>
</thead>
<tbody>
<tr>
<td>C7155: Primary Central Chondrosarcoma</td>
<td>C3737: Mesenchymal Chondrosarcoma</td>
</tr>
<tr>
<td>C5270: Cerbellar Papillary Meningioma</td>
<td>C3569: Malignant Cerbellar Neoplasm</td>
</tr>
<tr>
<td>C6430: Thymic Carcinoid Tumor</td>
<td>C3773: Neuroendocrine Carcinoma</td>
</tr>
<tr>
<td>C133894: Stage 0 Small Intestinal Adenocarcinoma</td>
<td>C7657: Intestinal Precancerous Condition</td>
</tr>
<tr>
<td>C39863: Adenocarcinoma of Skene Gland Origin</td>
<td>C6167: Urethral Adenocarcinoma</td>
</tr>
<tr>
<td>C15385: Excisional Biopsy</td>
<td>C64979: Diagnostic Surgical Procedure</td>
</tr>
<tr>
<td>C61145: Adenocarcinoma Cell with Eosinophilic Cytoplasm</td>
<td>C53644: Malignant Cell with Eosinophilic Cytoplasm</td>
</tr>
<tr>
<td>C121571: Leiomyosarcoma of Deep Soft Tissue</td>
<td>C9306: Soft Tissue Sarcoma</td>
</tr>
<tr>
<td>C6591: Peripheral Neuroblastoma</td>
<td>C4961: Malignant Peripheral Nervous System Neoplasm</td>
</tr>
<tr>
<td>C64000: Tubulostromal Adenoma of the Rat Ovary</td>
<td>C134942: Rat Neoplasms</td>
</tr>
<tr>
<td>C3758: Hepatocellular Adenoma</td>
<td>C36207: Digestive System Adenoma</td>
</tr>
<tr>
<td>C4826: Central Nervous System Neuroblastoma</td>
<td>C3568: Malignant Brain Neoplasm</td>
</tr>
<tr>
<td>C39951: Testicular Fibroma</td>
<td>C3709: Epithelial Neoplasm</td>
</tr>
<tr>
<td>C27404: Childhood Central Nervous System Mature Teratoma</td>
<td>C5591: Benign Childhood Central Nervous System Neoplasm</td>
</tr>
</tbody>
</table>

Table 3: Five examples of incorrect existing hierarchical relations obtained by our approach.

<table>
<thead>
<tr>
<th>Subconcept</th>
<th>Superconcept</th>
</tr>
</thead>
<tbody>
<tr>
<td>C66775: Borderline Ovarian Mucinous Adenofibroma</td>
<td>C4934: Benign Female Reproductive System Neoplasm</td>
</tr>
<tr>
<td>C33149: Muscularis Mucosa</td>
<td>C32209: Bladder Tissue</td>
</tr>
<tr>
<td>C4826: Central Nervous System Neuroblastoma</td>
<td>C3568: Malignant Brain Neoplasm</td>
</tr>
<tr>
<td>C38157: Metachronous Osteosarcoma</td>
<td>C4968: Secondary Malignant Neoplasm</td>
</tr>
<tr>
<td>C39951: Testicular Fibroma</td>
<td>C3709: Epithelial Neoplasm</td>
</tr>
</tbody>
</table>

Table 4: The precision of our approach in two ways to identify missing hierarchical relations based on the evaluation performed by the domain expert.

<table>
<thead>
<tr>
<th>Type</th>
<th># of suggested missing IS-As</th>
<th># of correct suggestions</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inheriting lexical attributes from ancestors within NLS</td>
<td>90</td>
<td>76</td>
<td>84.44%</td>
</tr>
<tr>
<td>Inheriting lexical attributes from all the ancestors</td>
<td>82</td>
<td>73</td>
<td>89.02%</td>
</tr>
</tbody>
</table>

Epithelium, but from the Stroma.

Next we give an example of the false positive cases which are not due to the existing erroneous hierarchical relations in NCIt. Our method suggests “C115093: Recurrent Oropharyngeal Undifferentiated Carcinoma” as a subtype of “C9268: Recurrent Malignant Nasopharyngeal Neoplasm” since it inherits lexical attribute “malignant”
from an ancestor “C150531: Recurrent Malignant Pharyngeal Neoplasm” and inherits lexical attribute “nasopharyngeal” from another ancestor “C4107: Nasopharyngeal Type Undifferentiated Carcinoma”. However, “C4107: Nasopharyngeal Type Undifferentiated Carcinoma” indicates that it looks like nasopharyngeal carcinoma under the microscope, but is not a nasopharyngeal carcinoma. Oropharyngeal carcinoma and nasopharyngeal carcinoma behave differently biologically, with nasopharyngeal carcinoma having a worse prognosis, and they are caused by different types of virus (HPV in oropharyngeal carcinoma, and EBV in nasopharyngeal carcinoma). Therefore, our suggestion is incorrect since our approach is incapable of capturing the subtle difference between “nasopharyngeal” and “nasopharyngeal type”.

5.2 Comparison with Previous Work

In our previous work\(^4\), we used six lexical patterns in NLSs to identify missing hierarchical relations in NCIt. One of the patterns was “Containment”, where we suggested hierarchical relations if the set of words of a concept is a subset of another. The “Containment” pattern was restricted to lower and upper bounds of the NLS while this work we have no such restriction. Also, we only considered the lexical attributes of the preferred term, while in this work we also enrich it with the lexical attributes of the ancestor terms. Furthermore, we perform three filtering steps to avoid obtaining incorrect suggestions.

The structural-lexical approach based on enriched lexical attributes was first introduced by Cui et al.\(^5\) to audit SNOMED CT. While our approach is similar to theirs, we perform a number of additional steps to improve performance and coverage. We do not skip considering an entire NLS if it contains stop words or antonym pairs as was done previously\(^5\). Rather, we perform a much fine-grained filtering by considering stop words and antonym pairs at the concept level, not the NLS level. Additionally, we also address an issue mentioned in Cui et al.’s work regarding incorrect suggestions when the set of words of a concept is a subset of another concept’s set of words. More importantly, in this work we introduce another way to enrich the lexical attributes of a concept: by considering all its ancestors (not only the ancestors within the NLS). This way was actually found to have a higher precision. Moreover, we do not put any restriction on the sizes of NLSs for evaluation in this work, while the evaluation was limited to small (size 4, 5, and 6) NLSs in the previous work.

5.3 Limitations and Future Work

While achieving a higher precision in suggesting missing hierarchical relations, our approach only covers a small portion of NLSs in NCIt (547 out of 9,512). New lexical patterns need to be identified to suggest remediations for the remaining unsolved NLSs. In this work, we enriched the lexical attributes of a concept by its ancestor lexical attributes. We expect to investigate into other methods that can be used for enriching, such as synonyms, definitions and other attribute relations. Another limitation of this work is that only one domain expert was involved in the evaluation. We plan to perform future evaluations by multiple domain experts to increase the robustness of the evaluation. In addition, although some of the failure suggestions of missing IS-A relations further revealed incorrect existing relations, it depended on the domain expert’s manual review. It would be desirable to develop automated methods to detect incorrect existing relations.

6 Conclusion

In this paper, we applied a structural-lexical auditing approach based on enriched lexical attributes of concepts in non-lattice subgraphs to suggest potential missing hierarchical relations in the National Cancer Institute thesaurus. This approach achieved a precision of 84.44% by inheriting lexical attributes from ancestors with NLSs, and a precision of 89.02% by inheriting lexical attributes from all the ancestors in the entire terminology, indicating the effectiveness of our approach. This approach could be generally applied to any biomedical terminology for quality assurance purposes.

Acknowledgment

This work was supported by the National Science Foundation (NSF) through grants IIS-1816805 and IIS-1931134, and the National Institutes of Health (NIH) National Cancer Institute through grant R21CA231904. The content is
solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.

References


Predicting Adverse Drug-Drug Interactions with Neural Embedding of Semantic Predications

Hannah A. Burkhardt, BS1, Devika Subramanian PhD2, Justin Mower, PhD2, Trevor Cohen, MBChB, PhD1
1University of Washington, Seattle, WA, 2 Rice University, Houston, TX

Abstract
The identification of drug-drug interactions (DDIs) is important for patient safety; yet, compared to other pharmacovigilance work, a limited amount of research has been conducted in this space. Recent work has successfully applied a method of deriving distributed representations from structured biomedical knowledge, known as Embedding of Semantic Predications (ESP), to the problem of predicting individual drug side effects. In the current paper we extend this work by applying ESP to the problem of predicting polypharmacy side effects for particular drug combinations, building on a recent reconceptualization of this problem as a network of drug nodes connected by side effect edges. We evaluate ESP embeddings derived from the resulting graph on a side-effect prediction task against a previously reported graph convolutional neural network approach, using the same data and evaluation methods. We demonstrate that ESP models perform better, while being faster to train, more re-usable, and significantly simpler.

Introduction
Drug-drug interactions (DDIs) are an important challenge to patient safety. Patients are commonly prescribed multiple drugs at a time, a phenomenon known as polypharmacy. The incidence of polypharmacy has been increasing1, and is generally very high in some patient groups, such as the elderly – the CDC reported that between 2011 and 2014, 66.8% of those 65 years and older take three or more prescription medications2. While multiple drugs can enhance each other’s efficacy and may be given in combination intentionally, many drug combinations can have unintended and harmful physiological consequences1. For example, if two medications affect the same biological pathway, they might compete for a substrate and the effect of one or both might be decreased or altered, side effects might be exacerbated, or new side effects might result from the drug interaction.

Identifying and accounting for polypharmacy effects is difficult1,3. While new drugs are tested for side effects during clinical trials, it is not feasible to investigate every possible combination of the new drug with all drugs currently on the market, let alone all investigational agents. As clinical trials do not exhaustively capture even the side effects of individual drugs, pharmacovigilance systems have been put in place for post-market surveillance to monitor for, and support regulatory action against, harmful drug effects. One such system is the Food and Drug Administration’s Adverse Event Reporting System (FAERS)4. Physicians, pharmacists, and others submit Adverse Drug Event (ADE) reports, which can then be analyzed for patterns of reporting, including identification of putative polypharmacy effects5. It is challenging to analyze these reports manually, so a range of data mining methods have been developed to detect drug/ADE associations6. Of importance for the current work, Tatonetti et al. developed a comprehensive database of DDI side effects, called TWOSIDES, using a data mining approach leveraging adverse event reports7.

Because of the combinatorial explosion of possible drug pairings to investigate, the development of computational methods to predict new polypharmacy side effects is a key task in pharmacovigilance. Several approaches to predicting drug-drug interactions have been described. Some approaches harness the idea that similar drugs will take part in similar drug interactions7. For this purpose, modeling of chemical and structural similarities has been explored7,8. Other approaches employ unstructured text, such as the Medline biomedical literature corpus, to extract interaction information with text mining techniques9. A third group of approaches to the DDI prediction problem uses binary classification techniques in order to classify a given drug pair as interacting or not interacting. A promising way to do this is to consider DDI prediction a link prediction problem, where drugs are nodes and side effects are the links between them6,10. However, predicting not only the presence of adverse effects of DDIs, but also the nature of such side effects, has not been extensively explored. For a detailed account of previously reported DDI prediction methods, the interested reader is referred to a recent comprehensive review by Vilar, Friedman and Hripcsak11.

The current work is inspired by Zitnik et al.’s graph convolutional network approach called Decagon11. Decagon makes use of the conceptualization of drug interactions as a graph, where drugs are nodes and side effects are edges, with different side effects being represented by edges of different types. Identifying polypharmacy side effects then becomes a multi-relational link prediction task. Graph embeddings are learned for drugs (nodes) as well as side effects.
(edges), and then used to predict which types of links are likely to exist between pairs of drugs. It is important to note that not only the presence of an interaction is predicted, but also the nature of this interaction, a distinguishing feature from other approaches such as Fokoue et al.’s link prediction model\(^\text{10}\). In fact, Decagon is reportedly the first approach to modeling different polypharmacy side effect types\(^\text{11}\). Decagon is trained on data from TWOSIDES, as well as various information sources for drug target and protein-protein interaction data, and achieves high performance as measured by area under the receiver operating curve (AUROC) and area under the precision-recall curve (AUPR) for the task of predicting side-effects given a drug pair. On this task, Decagon outperforms several baseline multi-relational link prediction approaches by at least 0.1 in mean AUROC across ADEs; RESCAL\(^\text{12}\), a multi-relational tensor factorization approach involving the decomposition of a matrix encoding drug-drug relationships into components representing drugs and side effects, achieves a mean AUROC of 0.693; and DeepWalk\(^\text{13}\), a procedure which produces neural embeddings based on a biased random walk through the node neighborhood within the network and then uses these embeddings as a basis for logistic regression, achieves a mean AUROC of 0.761. Additionally, Decagon is used to predict novel DDIs, a number of which were supported by the literature. However, this approach also has limitations. First, node embeddings are learned with a graph convolutional network\(^\text{14}\) approach in which d-dimensional node embedding vectors are linked by d-dimensional edge matrices. The computational demands of training such graph convolutional networks is high, limiting applicability to larger datasets – a problem that will not be eliminated by increasing compute power, as data set size increases at a similar rate\(^\text{15}\). Second, while the process produces vector embeddings for drugs (nodes), side-effect (edge) embeddings are represented by matrices and are thus challenging to re-use as they exist in a space of different dimensionality to the resulting drug vector space. Third, a simpler model might have advantages over Decagon because of its parsimony: if a smaller number of parameters is sufficient to model the underlying data, models with fewer parameters may avoid overfitting, and generalize better.

In the current work we propose an alternative approach to DDI prediction which builds upon Zitnik et al.’s graph conceptualization of the DDI prediction problem, but uses Embedding of Semantic Predications (ESP)\(^\text{16}\), which can address each of these challenges, for representation learning. ESP is a method for generating vector embeddings for biomedical concepts, such as drugs and side effects, from concept-relationship-concept triples called predications, using a neural network. ESP makes use of vector symbolic architectures, in which the composition (binding) and addition (superposition) operators are used to create transient embeddings for composite concepts (such as INHIBITS cytochrome p-450) from their component vectors. During training, embeddings are updated such that vectors for concepts occurring in predications with similar predicate-argument pairs become more similar (closer in vector space) to each other. Conversely, vectors for dissimilar concepts approach orthogonality in the vector space. The high dimensionality of vectors in ESP (e.g. 10,000 bits in binary vector implementations) makes it unlikely for vectors to be similar to each other by chance alone\(^\text{17}\) – an n-dimensional vector space contains \(2^n\) nearly orthogonal vectors\(^\text{18}\). Vectors are randomly initialized before the encoding process begins. For each positive triple in the training set, updates are made to the weight vectors representing the subject (S(s)), predicate (P(p)), and object (C(o)) as follows:

\[
S(s) + = P(P) \odot C(o) \times \alpha \times (1 - NNHD(S(s), P(P) \odot C(o)))
\]

\[
C(o) + = S(s) \odot P(P) \times \alpha \times (1 - NNHD(S(s) \odot P(P), C(o)))
\]

\[
P(P) + = S(s) \odot C(o) \times \alpha \times (1 - NNHD(S(s) \odot C(o), P(P))
\]

Note that, while ESP learns only one embedding for relationships (the \textit{predicate vector} P), two embeddings are learned for each concept: the \textit{semantic vector} (S) and the \textit{context vector} (C). \(\odot\) is the binding operator, \(\odot\) is the release operator (inverse of binding), and + is the superposition operator. The implementation of these operators depends on the vector symbolic architecture approach used. Here we use the Binary Spatter Code\(^\text{16,17}\), in which binding is the elementwise XOR. Because XOR is its own inverse, the release and binding operators are equivalent in this implementation. The superposition operator, which allows multiple vectors to be “added” together, retains the elementwise most common element (1 or 0) across the superposed vectors, with ties broken at random. The update steps are also modulated by a linearly decreasing learning rate (\(\alpha\)) and the similarity between the desired and current vectors, measured as the non-negative normalized Hamming Distance (NNHD)\(^\text{16}\). Analogous update steps are completed for negative samples, which are created by substituting random objects for the objects in positive triples (using -NNHD rather than 1-NNHD). The optimization objective, which is achieved through stochastic gradient descent, is the cross-entropy function defined as follows:

\[
\sum_{(s,P,o) \in D} \log \left( \text{NNHD}(S(s), P(P) \odot C(o)) \right) + \sum_{(s,P,\neg o) \in D^r} \log \left( 1 - \text{NNHD}(S(s), P(P) \odot C(\neg o)) \right)
\]
where \((s, P, ¬o)\) is a training triple and \((s, P, o)\) is a training triple with corrupted object (a negative example). In words, the optimization objective is to create representations of concepts that minimize the distance between subjects and representations of their corresponding predicate-object pairs, while maximizing the distance between subjects and representations of predicate-object pairs that they do not correspond to. An illustration of the ESP architecture used to learn the embeddings is shown in Figure 1.

**Figure 1.** Illustration of the ESP architecture used to generate 5-dimensional embeddings for a 10-concept vocabulary. Adapted from Cohen et al. 

This network trains matrices of weights consisting of the vector embeddings for each drug, activated separately during training using one-hot vector representations. For example, when encoding warfarin-BLEEDING-aspirin, the input and output vectors are one-hot vectors, with a 1 in the warfarin position and a 1 in the aspirin position, respectively. When multiplying the vector by the weight matrix, all columns are set to 0, except the ones pertaining to these particular drugs. Consequently, only the weight matrix components for the concepts and predicate of a triple are involved in the update step that encodes it. This update proceeds toward the optimization objective shown above. The weight matrices are updated so as to move \(S(\text{warfarin})\) closer to \(P(\text{bleeding}) \otimes C(\text{aspirin})\); \(C(\text{aspirin})\) closer to \(S(\text{warfarin}) \hat{} P(\text{bleeding})\); and \(P(\text{bleeding})\) closer to \(S(\text{warfarin}) \otimes C(\text{aspirin})\). We refer the interested reader to Cohen et al. for a detailed exposition of the algorithm and evaluations across a range of pairwise entity-level tasks.

Because similar representations will be learned for similar concepts, we can query the resulting vector space for concepts related to (similar to) a given concept (or concept composition). In our approach, we generate representations for drugs and side effects based on drug/side-effect/drug triples, resulting in input weights which serve as semantic vectors and output weights which serve as context vectors for drugs, and predicate vectors for side effects. In the generated vector space, the bound product of the semantic and context vectors for two drugs should then be similar to the predicate vectors of the polypharmacy side effects that might be observed when the two drugs are taken in combination. A query of interest might thus be: “What polypharmacy side effects might be expected when taking warfarin and aspirin together?” The response is the predicate vector most similar to the bound product of \(S(\text{warfarin})\) and \(C(\text{aspirin})\) over all possible side effects, minimizing \(\text{NNHD}(P(\text{side effect}), S(\text{warfarin}) \otimes C(\text{aspirin}))\).

ESP has previously been used successfully in predicting side effects for individual drugs. Most recently, Mower et al. utilized concept-relationship-concept triples from SemMedDB, a knowledge base extracted from the MEDLINE corpus of medical literature using SemRep, to construct an ESP model which provided input to a supervised machine learning model trained to predict which side effects might be seen for which drugs, achieving strong results with an area under the receiver operating curve (AUROC) of 0.96 and 0.95 across two different test sets. The success of ESP...
for predicting individual drug side effects raises the question of whether the approach might be used to predict DDIs as well. In this paper, we explore the hypothesis that ESP has utility for the polypharmacy side effect prediction task.

**Methods**

When selecting data for pharmacovigilance research, balancing level of curation with representation of emergent signals (such as those that might be found in ADE reports) is challenging\(^1\). As a result, prior evaluations of pharmacovigilance approaches have used a range of diverging data sets for training and evaluation. Because of the large variety of available data and their particular biases and limitations, it can be difficult to directly compare the performance of pharmacovigilance models\(^1\). To minimize this problem, we elected to use the same dataset used for evaluation of a recently published approach to identifying specific drug-drug adverse effects, to facilitate direct comparison of our results. Commendably, Zitnik et al. published both the data sets used in their approach and their code. As a result, we were able to use the same data as well as the same train/test splits, as produced by the publicly available code. The data were downloaded from Stanford’s SNAP website\(^2\) and the code was downloaded from Marinka Zitnik’s GitHub page\(^3\). We performed the additional preprocessing steps described by Zitnik et al., taking care to follow their methods as closely as possible. This included discarding any triples for side effects that occurred in fewer than 500 drug interaction triples, reducing the total number of side effects in the dataset from 1317 to 963.

After modifying the Decagon code to output the testing set as well as the data subset used for training, we allowed Decagon to complete several epochs of training. However, in our own execution of the protocol, we were not able to run Decagon for up to 100 epochs as described by Zitnik et al. (with documented early stopping) within a reasonable time frame, possibly due to differences between our groups’ computational resources. With a single NVIDIA Tesla p40 GPU we observed training times of roughly 36 hours per epoch, with a one-time up-front setup and initialization time of about 6 hours. Consequently, we obtained predictions over the testing set after allowing Decagon to train for 12 epochs only. We present both these results and the previously-published Decagon results showing best-documented performance, enabling fair comparison within constraints of our timeframe and computational resources.

The downloaded data set consists of 9,643,506 triples, of which 7,323,790 are drug-drug interaction triples, 2,289,960 are protein-protein interaction pairs, and 29,756 are drug target pairs (linking drugs to proteins). While the downloaded data set also contained individual drug side effects, these data are not used in the published code and were consequently left out for the purposes of the current research as well. The training data consist of 80% of the total data, with the remaining 20% being equally divided between the validation and test sets. In Decagon, the validation set is used to determine the early stopping point by computing the cross-entropy validation loss over the validation triples after each epoch; training is stopped early if validation loss does not improve for 2 consecutive epochs. Testing is performed exclusively over the testing set triples. It is of note that in the original Decagon work, the authors doubled the size of the dataset in the downloadable data files by also including all training triples in their inverted form, treating the “reverse” side effect edge types as separate edge types, resulting in a total of 963*2=1926 side effects (drug-drug edge types). Later, for purposes of analysis, we examine AUCs per side effect, only considering the original 963 side effects.

From these data, we created subject-predicate-object triples (predications) as input for ESP training (Table 1).

<table>
<thead>
<tr>
<th>Data type</th>
<th>Format</th>
<th>Raw data example</th>
<th>Predication example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug-drug interaction</td>
<td>drug1, drug2, side effect</td>
<td>aspirin, warfarin, kidney failure</td>
<td>aspirin KIDNEY_FAILURE warfarin</td>
</tr>
<tr>
<td>Protein-protein interaction</td>
<td>protein, protein</td>
<td>GPRIN1, GJB7</td>
<td>GPRIN1 INTERACTS_WITH GJB7</td>
</tr>
<tr>
<td>Drug target</td>
<td>drug, protein</td>
<td>aspirin, apoTF</td>
<td>aspirin TARGETS apoTF</td>
</tr>
</tbody>
</table>

An ESP model was trained using the open-source Semantic Vectors package\(^{15,24,25}\) pre-release version 5.9 with vector dimensionality of 16,000 bits. Unlike previous releases, this version allows predicate vectors to be trained (in addition to semantic and context vectors). This capability is essential for our approach to DDI prediction, as we generate trained embeddings for side effects (represented as predicates), as well as drugs.

We converted the testing data set to a list of similarity comparisons to be made. For each triple in the test set, we bind the semantic vector for the first drug with the context vector for the second drug and compare the similarity of the resulting compositional vector with the predicate vector for the side effect:

\[
NNHD(P(\text{side effect}), S(\text{drug 1}) \otimes C(\text{drug 2}))
\]
The result is a similarity score between 0 and 1. Identical vectors score 1 and orthogonal vectors 0. It is important to note here that the publicly available Decagon data sets, like most drug-drug interaction data sets, only contain positive examples, i.e. data points with a ground truth label of 1. ESP originates from methods that learn representations of words from free text, where the same problem exists; ESP therefore automatically performs negative sampling in the training process to account for this. However, during testing of ESP, we must test using generated negative examples. The published Decagon code generates negative examples by choosing 2 random drugs and connecting them with a random side effect, making sure that the resulting triple (edge) is not observed in the original, positive data. Additionally, for each side effect, we generate the same number of negative examples as we have positive examples. The negative testing examples produced by the Decagon code were used for evaluating the ESP model.

Evaluation was performed with Anaconda® version 4.6.8 and Python version 3.6.8 using the scikit-learn package version 0.19.1. As in the original Decagon protocol, AUROC and AUPRC were calculated on a per side effect basis. Precision at $k$ was also calculated, with $k=50$, and all three metrics were averaged across all side-effects.

While 16,000-bit vectors have been shown to perform well on a range of prior ESP evaluations, training time and model parsimony might be improved by using lower dimensional vectors. In order to determine the minimal space requirements of a model with acceptable accuracy, we analyzed several lower embedding dimensionalities. Due to the distributed nature of the vector representations (information is spread as a pattern across the entire vector), it is possible to test the loss in accuracy for smaller dimensional vectors without re-training, but rather by truncating the trained vectors. We thus truncated the trained 16,000-dimensional representations to 8000, 4032, 2048, 1024, 512, 256, 128, and 64 bits and computed performance metrics for each of the resulting vector spaces.

Querying the vector space for familiar concepts presents an opportunity to qualitatively evaluate whether the learned representations are meaningful. Using the Semantic Vectors open source software, the vector space was queried using various drugs, side effects, and compositional concepts to assess whether the space is intuitively interpretable.

Finally, we investigated the proximity of related side effects to each other using a visualization of the side effect vectors, created using Uniform Manifold Approximation and Projection (umap-learn package14,27, version 0.3.7), which projects multi-dimensional representations into two dimensions. Individual side effects were colored according to their class as listed in the downloaded Decagon datasets. Many side effects did not belong to any category, and some categories had few or no side effects in them. Of the remaining categories, we visualized the 6 most frequently occurring side effect classes.

Code and data for all preprocessing, training, and evaluation steps are available on Github28,29.

**Results**

The performance metrics for ESP and Decagon, averaged over 963 side effects, are shown in Table 2. The mean AUROC for the 8-epoch ESP model was 0.903 (range 0.841-0.977); the mean AUPRC was 0.875 (range 0.779-0.976); and the mean AP@50 was 0.865 (range 0.550-1.0). These results exceed both the published Decagon results (in Table 2) and those of our locally-trained Decagon model after 8 epochs: mean AUROC of 0.855 (range 0.259-0.949); a mean AUPRC of 0.793 (range 0.366-0.934); and a mean AP@50 of 0.638 (range 0.066-0.950).

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean AUROC</th>
<th>Mean AUPRC</th>
<th>Mean AP@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESP (4 epochs)</td>
<td>0.896</td>
<td>0.868</td>
<td>0.861</td>
</tr>
<tr>
<td>ESP (8 epochs)</td>
<td><strong>0.903</strong></td>
<td><strong>0.875</strong></td>
<td><strong>0.865</strong></td>
</tr>
<tr>
<td>Decagon (4 epochs)</td>
<td>0.826</td>
<td>0.768</td>
<td>0.644</td>
</tr>
<tr>
<td>Decagon (8 epochs)</td>
<td>0.855</td>
<td>0.793</td>
<td>0.638</td>
</tr>
<tr>
<td>Decagon (published)</td>
<td>0.872</td>
<td>0.832</td>
<td>0.803</td>
</tr>
</tbody>
</table>

Table 2. Performance metrics for ESP and Decagon.

Our 4-epoch ESP model took roughly 3.5 hours to train with 16,000 dimensions (bits), including setup time and time spent making predictions over the test set. Each epoch took an average of 50 min. 4 epochs of training Decagon were completed in 6 days and 4 hours including setup and test set prediction, with an average of 36 hours per epoch.

The mean area under the receiver operating characteristic curve over 963 side effects for different vector dimensionalities, created by truncating the 16,000-dimensional vectors, are shown in Figure 2.
Example queries of the trained vector space are shown in Table 3. The search returns the vectors that most similar to the query term, along with a similarity score.

Table 3. Example searches of the predicate vector space. S,C,P: Semantic, Context and Predicate Vectors

<table>
<thead>
<tr>
<th>Cue</th>
<th>Explanation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(KIDNEY_FAILURE)</td>
<td>What side effects occur in similar drug combinations as kidney failure?</td>
<td>1.000:KIDNEY_FAILURE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.904:ACUTE_KIDNEY_FAILURE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.871:ANAEMIA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.870:CARDIAC_FAILURE</td>
</tr>
<tr>
<td>S(aspirin)⊗C(warfarin)</td>
<td>What side effects might be caused by taking aspirin and warfarin together?</td>
<td>0.363:FIBROSING_ALVEOLITIS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.347:PAROTITIS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.341:HAEMARTHROSIS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.339:NECK_MASS</td>
</tr>
</tbody>
</table>

A visualization of how vectors for side effects that are similar to each other are located with respect to one another within the learned embedding space is shown in Figure 3.

Figure 3. UMAP clustering of a subset of side effect groups. Categories are from Decagon's published data sets.

Discussion

ESP outperforms Decagon (both best published performance and our locally-trained model) when predicting polypharmacy side effects by all measures. Furthermore, ESP achieves this performance while training a factor of 43.2 times faster per epoch than Decagon, and for 8 epochs only. The best performing Decagon models were run up to 100 epochs (the number of epochs prior to early stopping was unfortunately not documented). We restricted our locally-trained model to a maximum of 8 epochs on account of resource constraints, which explains the modest difference in performance from previously published results. Additionally, we note that ESP produces re-usable embeddings of the same dimensionality drugs and side effects, which can be used to answer various questions using vector symbolic architecture operations. Using this approach, as well as projection visualization using UMAP, we were able to interpret the embeddings produced by the model.
The high performance that ESP was able to achieve is notable. Considering that Decagon substantially outperformed several baseline approaches, the improved performance of ESP over Decagon puts this model among the top in class. The multimodal network representation and modeling of multiple side effects simultaneously, which allows sharing of model parameters across side effects, is an advantage of Decagon over some alternative polypharmacy side effect prediction models that is shared by ESP, and may be in part responsible for improved performance as the model can learn to generalize across similar concepts (such as ‘kidney failure’ and ‘acute kidney failure’, shown in Table 1).

While training time may not be the most significant factor in determining which models are superior when abundant computational resources are available, a faster model is preferable if degradation of accuracy is not a tradeoff that has to be made. Less extensive hardware requirements enable academic researchers with relatively modest computational resources to reproduce and build upon extant research, thus enabling scientific progress through improved equity. But even for those research groups with high-end computing clusters and multiple GPUs at their disposal, it should be considered that training complex neural network models can incur significant energy costs, with carbon footprints equivalent to as much as five times that of an average car, including manufacture\(^{30}\). In comparison, ESP can be trained on commodity laptops or servers. Financial and environmental costs aside, one may argue that resource intensive computations will become progressively easier as processing power increases in accordance with Moore’s law (at least for some time\(^{11}\)); however, this is a difficult case to make in informatics, where much work is based on data sources that increase in size at an incredible pace themselves\(^{14}\). As we harness biomedical literature, electronic health record data, social media sources, and adverse event reports, we also observe their sustained and sometimes explosive (e.g. EHR and patient reported data) growth. Currently, DDI prediction models should be retrained whenever new adverse event report data become available (e.g. weekly or monthly). With the advent of large volumes of patient-level data, DDI prediction models have the potential to be personalized to individual patients’ demographics, health status, and genetic makeup, requiring scores of different models to be trained and re-trained, further increasing training costs. ESP conforms to previously formulated recommendations to prioritize development of efficient models\(^{30}\) and can help promote research equity by allowing the use of commodity hardware.

The ability to learn representations that correlate with human judgment of similarity between concepts is an advantage of ESP\(^{16}\). In this work, we were able to demonstrate that embeddings learned for related side effects, such as kidney failure and acute kidney failure, are similar (Table 3). However, other types of organ failure are also retrieved by this search. A possible explanation is that patients with multi-organ failure appear in FAERS, and so different sorts of organ failure may become associated with one another. If these organ failures are attached to reports containing the same drug pairs (e.g. drugs used in sepsis), a statistical model will find associations between these drugs and every sort of organ failure. The similarity-based search for side effects using the bound product S(aspirin)⊗C(warfarin) yielded, among others, alveolitis, parotitis, and hemarthrosis. Development of lung disease has previously been reported to be associated with aspirin treatment\(^{32}\); additionally, warfarin has been characterized as a drug causing salivary gland swelling and pain\(^{31}\). Aspirin’s and warfarin’s possible roles, respectively, in exacerbating these adverse events has yet to be established. We also found similarity of S(aspirin)⊗C(warfarin) with hemorrhosis, a bleeding disorder, which is reasonable, considering that warfarin and aspirin are both anticoagulants. We note that similarity search results may refer to indications rather than side effects. Even though Tatonetti et al.\(^{5}\) applied corrections for confounding when creating the TWOSIDES set, with the goal that indications would not appear as side effects, it appears that this effect may not have been entirely eliminated, resulting in our model learning from indication-related statistical patterns remaining in this dataset.

Not all side effect representations for a given side effect category clustered together well in our projection visualization. For example, side effects in the reproductive system, cognitive disorder, and hematopoietic system effect classes are well separated from each other, but nervous system side effects are found throughout the vector space. We suspect that one contributing factor for this is the limited number of types of data in the training set. The only information that ESP has an opportunity to encode into the concept embeddings are polypharmacy side effects and drug target information. As discussed above, side effects appear to cluster because they are actually indications common to particular drug combinations – an additional complicating factor. This was also discussed by Zitnik et al., who reported that many side effects tend to co-occur in this particular data set\(^{11}\). As a result, co-location of embeddings in space is not as interpretable as it might be for richer data sets, with more concept and relation types, which might contribute to more meaningful similarities between the various side effects being learned. Previous applications of ESP have involved rich data sets with multiple relationship types and a number of different types of concepts, which resulted in more readily interpretable similarities.

The parsimony of the ESP model is a desirable feature. While Decagon utilizes approximately 87 million trainable parameters, which are each 32-bit floating point numbers – a total of approximately 2.8 billion bits of training
parameters – using 16,000-dimensional vectors, ESP performs better with only about 36 million single bit parameters, which is over 77 times fewer bits of representational capacity. In fact, even with 1,024 dimensions, ESP approximates Decagon performance, and this dimensionality corresponds to only 2 million trainable bits of representational power – over three orders of magnitude less than Decagon. At 64 bits per embedding – a mere 144,000 bits of trainable parameters – ESP achieves performance that is still higher than the baseline approaches documented by Zitnik et al. This suggests that the vast representational capacity of a trained Decagon model is unnecessary for the task of predicting DDIs that appear in TWOSIDES. More stringent evaluations may be required to reveal any advantages in performance that its larger capacity and more computationally demanding training procedure may confer.

This work has several limitations. Firstly, it is important to note that the data used for training may poorly approximate ground truth. TWOSIDES is a database of information mined from text sources. While a small number of novel drug interactions reported in TWOSIDES have been corroborated through investigation of relevant patient records as well as laboratory experiments, the database contains many unvalidated associations, including some unintuitive side effects that may not withstand curation, such as ‘Mumps’ or ‘Fracture’. These side effects do give us an idea of what the drug interactions may be; for instance, if mumps is increasingly observed in patients taking a certain combination of drugs, it may be plausible to infer the polypharmacy side effect to be a weakening of the immune system. Similarly, if fractures have increased prevalence in the patient population taking particular drugs in combination, we may consider the idea that the interaction of the drugs causes decreased bone density. Yet, using a curated reference set of drug interaction data is preferable over using data obtained through data mining. As discussed above, making models directly comparable is a challenge, which is why we chose to use the same datasets as a comparable previous approach.

Another limitation of this approach is that directionality of relationships between concepts is not encoded. DDIs may be considered symmetric relationships between drugs, as drug A interacting with drug B would also mean that drug B interacts with drug A. Here, we reverse and append all DDI triples in the dataset during preprocessing, creating a separate side effect type representing the “reverse relationship”, as Zitnik et al. did. In other words, for each edge (with a side effect specific edge type) pointing from drug A to drug B, we also add an edge from drug B to drug A, which has a separate, new edge type. While the additional triples (edges) benefit the training of drug embeddings, the predicate embeddings to which they contribute (redundant side effect representations) are not used when making predictions. Conversely, the inverse relationship, despite being known to be true, is not encoded in the retained side effect embedding. Similarly, drug target triples are reversed and appended to the dataset with a new edge type. In the drug target case, this matches the directionality of the relationship: if drug A targets protein B, we cannot also say that protein B targets drug A. Directionality would be obscured if we reversed and appended the drug target triples with the same edge type. Finally, protein-protein interaction (PPI) triples, which are inherently symmetric, are treated differently. Decagon quadruples all PPI triples, first by reversing the order and then by creating an alternative relationship, such that the nodes for protein A and protein B are connected by a total of 4 edges: 2 “interacts_with” edges, one pointing from A to B and one from B to A, and 2 “reverse_interacts_with” edges, one pointing from A to B and one pointing from B to A. While ESP can encode directionality directly, we did not explore this here and instead chose to follow the same input data preprocessing protocol as Zitnik et al. in order to facilitate direct comparability.

In this research, we have followed the training and evaluation protocol reported by an earlier study as closely as possible for the purpose of comparative evaluation, leveraging both written descriptions of the protocol and publicly available data and software. However, during this process, we found some details in the reported protocol ambiguous, and found apparent inconsistencies between the approach described in the paper and the code it refers to. For instance, the number of side effects and proteins reported by Zitnik et al. does not exactly match the number we found in the data ourselves (963 side effects with 500 or more drug pairs, compared to the reported 964; 19081 unique drug targets compared to the reported 19085). Also, neither cross validation nor early stopping are implemented in the publicly released code, though both methods were reported in the paper. The 10% of data reserved as validation set consequently went completely unused in the published code. As a result of the computational demands of training Decagon, a test of statistical significance, which would require Decagon and ESP to each be trained repeatedly, was not performed. Additionally, we found that all data were split into training and test sets, including protein-protein interaction and drug-target data, even though we are not concerned with making predictions for these data subsets. This equates to discarding a portion of the protein-protein and drug-target data without good reason, other than allowing a simple 80/10/10 split across all types of triples. While these discrepancies were reconciled by our group to our best effort, subtle differences in our evaluation pipelines may not have been totally eliminated.

This project has established a foundation for future work. Having demonstrated the accuracy of ESP as compared to previously reported models while keeping the input data constant, we plan on applying the technique to datasets that are (at least partially) curated rather than statistically derived, e.g. the interaction designations in FAERS reports or
the reference set published by Ayvaz et al.\textsuperscript{35}. While accuracy measures such as AUROC and AUPRC may not be affected by this change, the resulting model would presumably have more validity when applied to real data.

Incorporating additional knowledge may also present an opportunity to improve this modeling approach. ESP is a method for encoding a knowledge graph; thus, we can generate embeddings for any number of concepts and relationships based on their connections to each other. Many sources of knowledge regarding drugs and their interactions may be relevant. For example, drug classifications, molecular structure information, and additional information about drug target molecules may allow us to train more comprehensive embeddings which could even be applied to new drugs prior to the appearance of any adverse drug event reports. On the patient level, ESP models for predicting DDIs may help make strides in precision medicine, e.g. by incorporating patients’ individual genetic profiles in the model. Training patient specific models would be made feasible in part due to ESP’s efficiency in both training time and model complexity, each being only about 1-2% of what previously reported models require.

Conclusion

Many individuals are at risk of experiencing adverse events caused by drug-drug interactions. Because DDIs are poorly characterized before drugs are released to market, efficient and accurate analysis of post-market surveillance data is critical to patient safety. ESP, a representation learning method with demonstrated utility in the pharmacovigilance space, provides an alternative approach to DDI prediction. In this work, we have shown that application of ESP to adverse drug event data can yield accurate predictions of polypharmacy side effects, with performance exceeding that of the best reported machine learning models on this task. Being considerably more lightweight in terms of required compute power and model complexity than previously published approaches, ESP can be expected to scale up without imposing undue financial and ethical burdens on researchers, particularly as the volume of data, number of models to train, and frequency of re-training increase. Additionally, we have demonstrated that ESP produces meaningful representations of biomedical concepts that can enable biomedical discovery in novel ways, providing insights that pave the way for further development of ESP models for pharmacovigilance and other applications.

Acknowledgments

This work is supported by a training fellowship (Grant No. 3T15 LM007442, Biomedical and Health Informatics Training Program) and U.S. National Library of Medicine Grant (R01LM011563), Robust Inference from Observational Data with Distributed Representations of Conceptual Relations.

References

Comparison of Prediction Model Performance Updating Protocols: 
Using a Data-Driven Testing Procedure to Guide Updating

Sharon E. Davis, MS¹, Robert A. Greevy, PhD¹, Thomas A. Lasko, MD PhD¹, 
Colin G. Walsh MD MA¹, and Michael E. Matheny, MD MS MPH¹,²
¹Vanderbilt University School of Medicine, Nashville, TN; ²VA Tennessee Valley 
Healthcare System, Nashville, TN

Abstract

In evolving clinical environments, the accuracy of prediction models deteriorates over time. Guidance on the design of model updating policies is limited, and there is limited exploration of the impact of different policies on future model performance and across different model types. We implemented a new data-driven updating strategy based on a nonparametric testing procedure and compared this strategy to two baseline approaches in which models are never updated or fully refit annually. The test-based strategy generally recommended intermittent recalibration and delivered more highly calibrated predictions than either of the baseline strategies. The test-based strategy highlighted differences in the updating requirements between logistic regression, L1-regularized logistic regression, random forest, and neural network models, both in terms of the extent and timing of updates. These findings underscore the potential improvements in using a data-driven maintenance approach over “one-size fits all” to sustain more stable and accurate model performance over time.

Introduction

Clinical environments are continuously evolving through changes in patient case mix, outcome prevalence, and clinical practice shift. These changes can cause the performance characteristics of clinical prediction models to deteriorate over time. As prediction models are increasingly incorporated into electronic health records to support decision-making by providers and patients, model updating strategies to sustain performance are becoming critical components of model implementations. Despite literature documenting performance drift and the availability of multiple updating methods, guidance on the design of model maintenance plans is limited and provides little insight into how differences between learning algorithms may impact updating requirements.

While a lack of model updating can harm the performance and utility of predictions, common model maintenance strategies, such as regularly scheduled model refitting, may be inefficient or even detrimental. The assumption that a new model is necessary neglects information gleaned from previous modeling efforts and can lead to overfit models that lack generalizability, especially when updating datasets are smaller than development cohorts. Such pre-defined updating plans also fail to account for variations in the response of different modeling methods to changes in clinical environments, which may impact the extent and form of shifts in model accuracy. Alternatively, a range of recalibration methods are available that may correct deteriorations in performance by incorporating information in recent observations while retaining information in existing models and reducing the risk of overfitting. Recalibration may thus be more appropriate than refitting for models in clinical use when equivalent or improved performance can be achieved by the former.

Recently proposed methods seek to provide data-driven guidance on when to retain the current model, apply recalibration of varying degrees of complexity, or refit a model. We described a nonparametric testing procedure to select between competing updating methods while minimizing overfitting and taking the updating sample size into consideration. As a baseline, the testing methods under consideration are retention of the existing model, intercept correction, linear logistic recalibration, flexible logistic recalibration, and model refitting. The testing procedure is designed to recommend the simplest updating method that does not compromise accuracy that may be achievable through more complex adjustments. In contrast with other tests, our procedure is customizable and widely applicable to models for categorical outcomes regardless of the underlying learning algorithm.

In this study, we explore whether the long-term performance of clinical prediction models is improved through a data-driven approach to model maintenance. We compare three updating strategies—retention of the original model, predefined model refitting, and application of recommendations of the nonparametric testing procedure. These updating strategies are applied to a model for 30-day mortality after hospital admission in a national population of veterans for which calibration drift, and variability in drift by modeling methods, has been documented across multiple years. We
assess differences in discrimination and calibration over time under each updating strategy, as well as whether and how the learning algorithm underlying the model impacts updating requirements and accuracy.

**Methods**

We developed models for 30-day mortality after hospital admission among patients admitted to Department of Veterans Affairs (VA) facilities nationwide using logistic regression (LR), L1-regularized logistic regression (L1), random forests (RF), and neural networks (NN). Model predictors and cohort eligibility criteria have been detailed in previous work. Each model was developed using a common set of predictors and admissions occurring in 2006. Data from 2007-2013 were collected for both updating and validation. Previous work indicated the LR and L1 were most subject to calibration drift in this population while the NN model did not experience significant calibration drift. Updating was undertaken for all models on an annual basis at the end of 2007 through 2012, with updates based on admissions in the prior 12 months and applied to admissions in the following 12 months. Admissions occurring in 2013 served as validation data for updates at the end of 2012.

We implemented three competing strategies to update the LR, L1, RF, and NN models over time in yearly increments. As a baseline, we retained the original models developed on 2006 admissions and applied these models to all subsequent admissions through 2013. The second updating strategy called for annually refitting each model using all admissions that accrued over the prior 12 months. Hyperparameters for the L1, RF, and NN models were tuned annually using 5-fold cross-validation. Admissions in each year were assigned predicted probabilities based on the prior year’s models. The third updating strategy selected the updating approach for each model based on a nonparametric testing procedure that we have developed for this purpose. A simplified illustration of the testing procedure is presented in Figure 1. The procedure selects between the retention of the current model, intercept correction, linear logistic recalibration, flexible logistic recalibration, or model refitting. Users may specify additional updating approaches for consideration as desired. A two-stage bootstrapping framework accounts for both overfitting and sample size, while avoiding assumptions about the structure of the model’s learning algorithm. In the first bootstrapping stage, updated predictions on the out-of-sample observations are stored to construct a population of pooled holdout predictions. These observations are leveraged by the second bootstrapping stage to characterize the performance of each updating approach. Paired differences in a user-specified performance metric, in this case the Brier score, between the best performance updating approach and simpler approaches are evaluated in the decision stage, with the procedure recommending the simplest update exhibiting no statistically significant difference in accuracy. We specified a Type I error rate of 0.05 for these comparisons.

![Figure 1](image-url)

**Figure 1.** Simplified illustration of the nonparametric testing procedure used to select updating methods under the data-driven annual updating strategy. Hexagons indicate processes; canisters indicate datasets.

Updating sequences were retained over multiple years as needed, allowing updates to build on any prior adjustments to the model. For example, a model based initially on Year 0 admissions was applied to Year 1 admissions. At the end of Year 1, the testing procedure recommended either continued use of the existing model, adjustment of the existing model through recalibration, or replacement of the model with a newly refit model. This updated version of the model was used to generate predictions for Year 2. Following Year 2, the testing procedure considered whether any additional updates to the model as adjusted after Year 1 were warranted, not whether to adjust the original Year 0 model. If additional updating was recommended by the test, those changes were applied in addition to the existing Year 1
adjustments. At any point, if the test recommended refitting the model, then all previous models and sequences of adjustments were replaced by a new model moving forward.

We conducted analyses to assess the influence of each updating strategies on the long-term performance of the LR, L1, RF, and NN models, as well as to compare the updating requirements of the different models. Models were assessed for both discrimination (area under the receiver operating curve, AUC) and calibration (calibration curves, observed to expected outcome ratio, Cox intercept and slope, and estimated calibration index)\textsuperscript{15-17}. We evaluated overall performance of the four models under each updating strategy across the entire validation and updating period (2007-2013). We further measured performance over time under each strategy on a monthly basis. In order to identify any differences in updating requirements, we recorded the recommendations of our nonparametric testing procedure for each of the four modeling methods.

This study was approved by the Institutional Review Board and the Research and Development committee of the Tennessee Valley Healthcare System VA.

Results

This study included data on 1,893,284 admissions to VA facilities nationwide. The initial 2006 models were developed on 235,548 admissions. The validation and updating set, consisting of admissions from 2007 through 2013, included 1,657,736 admissions with a mean of 236,819 per year. Fewer admissions were captured in 2013 because December admissions lacked sufficient follow-up time to ascertain the outcome. The overall 30-day mortality rate was 4.9%. This outcome rate was stable over time, with annual mortality varying between 4.7% and 5.0%.

<table>
<thead>
<tr>
<th>Year</th>
<th>Update set</th>
<th>LR</th>
<th>L1</th>
<th>NN</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>Flexible logistic recalibration</td>
<td>Flexible logistic recalibration</td>
<td>Refit</td>
<td>Linear logistic recalibration</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>No change</td>
<td>No change</td>
<td>Refit</td>
<td>No change</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Intercept correction</td>
<td>No change</td>
<td>Refit</td>
<td>Linear logistic recalibration</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>No change</td>
<td>No change</td>
<td>Refit</td>
<td>No change</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Intercept correction</td>
<td>Intercept correction</td>
<td>Refit</td>
<td>No change</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>No change</td>
<td>No change</td>
<td>Refit</td>
<td>No change</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Annual updating recommendations for each modeling approach

Test-based recommendations for annual updates of the four models are noted in Table 1. The nonparametric testing procedure recommended model refitting each year for the NN model. The testing procedure recommended flexible logistic recalibration after the first year for the LR and L1 models, as well as linear logistic recalibration for the RF model. For the L1 model, the recalibration adjustments incorporated after the first year were maintained until the 5th year after model development, at which point an additional intercept correction was recommended. The RF model was also updated again after the 3rd year. Continued periodic updating across the study period was recommended for the LR model, with the testing procedure recommending some degree of recalibration every other year.

Performance of the four models under each updating strategy over the entire validation and updating period (2007-2013) is reported in Table 2 and calibration curves are presented in Figure 2. In most cases, discrimination was unchanged by updating, the exception being the NN model for which refitting (and the test-based strategy) increased the AUC from 0.77 to 0.80. Annually refitting the models improved calibration across the study period compared to
the original models (p<0.05). However, for all models, predictions based on test recommendations exhibited improved calibration compared to predictions based on either the original model without updating or the annually refit models (p<0.05)—the only exception being the NN model for the test-based strategy reduced to refitting. Differences in calibration across updating strategies were highlighted by the calibration curves and most apparent when focusing on the lower risk portion of the curves where over 95% of observations occur. For the LR model, the calibration curves under the test-based updating strategy captured more of the ideal 45° calibration line than either the annual refit or original models. For the RF model, none of the updating strategies resulted in calibration across a large range of probabilities and the calibration curves of all three strategies follow similar patterns; however, both the refitting and test-based updating strategies moved the calibration curve closer to the ideal 45° calibration line for the risk range where most observations fell. In the densely populated risk range, although the magnitude of miscalibration of the L1 model was similar between the refitting and test-based updating strategies, the refitting approach erred toward underprediction while the test-based strategy erred toward overprediction.

### Table 2. Overall performance by modeling approach and annual updating strategy

<table>
<thead>
<tr>
<th>Model</th>
<th>Updating Strategy</th>
<th>AUC</th>
<th>O:E</th>
<th>Cox Intercept</th>
<th>Cox Slope</th>
<th>ECI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>No updating</td>
<td>0.849</td>
<td>0.876</td>
<td>-0.227</td>
<td>0.970</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>Refitting</td>
<td>0.850</td>
<td>0.981</td>
<td>-0.052</td>
<td>0.987</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>Test-based</td>
<td>0.849</td>
<td>0.953</td>
<td>-0.073</td>
<td>0.994</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.847, 0.850]</td>
<td>[0.871, 0.882]</td>
<td>[-0.242, -0.214]</td>
<td>[0.964, 0.976]</td>
<td>[0.027, 0.032]</td>
</tr>
<tr>
<td>L1</td>
<td>No updating</td>
<td>0.846</td>
<td>0.815</td>
<td>-0.221</td>
<td>1.014</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>Refitting</td>
<td>0.846</td>
<td>0.936</td>
<td>0.005</td>
<td>1.038</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Test-based</td>
<td>0.846</td>
<td>0.937</td>
<td>-0.081</td>
<td>0.999</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.845, 0.847]</td>
<td>[0.810, 0.821]</td>
<td>[-0.237, -0.207]</td>
<td>[1.008, 1.021]</td>
<td>[0.036, 0.041]</td>
</tr>
<tr>
<td>RF</td>
<td>No updating</td>
<td>0.837</td>
<td>0.842</td>
<td>-0.031</td>
<td>1.080</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>Refitting</td>
<td>0.837</td>
<td>0.950</td>
<td>0.127</td>
<td>1.082</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>Test-based</td>
<td>0.837</td>
<td>0.939</td>
<td>-0.035</td>
<td>1.017</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.836, 0.838]</td>
<td>[0.943, 0.956]</td>
<td>[0.096, 0.153]</td>
<td>[1.070, 1.094]</td>
<td>[0.024, 0.028]</td>
</tr>
<tr>
<td>NN</td>
<td>No updating</td>
<td>0.770</td>
<td>0.914</td>
<td>-0.187</td>
<td>0.965</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>Refitting</td>
<td>0.800</td>
<td>0.991</td>
<td>-0.104</td>
<td>0.961</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Test-based</td>
<td>0.800</td>
<td>0.991</td>
<td>-0.104</td>
<td>0.961</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.798, 0.802]</td>
<td>[0.984, 0.997]</td>
<td>[-0.122, -0.087]</td>
<td>[0.955, 0.967]</td>
<td>[0.003, 0.004]</td>
</tr>
</tbody>
</table>

Figure 3 displays monthly calibration of all four models under each updating strategy using the estimated calibration index (ECI). This stringent measure of calibration decreases toward 0 as calibration improves17,18. Without updating, calibration of the LR, L1, and RF models decayed over time. Both refitting and test-based updates improved calibration compared to the original model over the course of the 7 years following initial model development. The RF model was an exception to this pattern. In some time periods, such as 2009-2010, refitting the RF model each year did not improve performance compared to the original model without updating. Although calibration of the NN model was stable over time compared to the other models, annually refitting the NN model improved calibration and reduced month-to-month variability in performance. With the exception of the NN model, monthly ECIs under the test-based updating strategy were generally lower and less variable compared to ECIs under the refitting strategy. For those points at which the nonparametric testing procedure recommended updating, ECIs over the prior 12 months (i.e., performance among those admissions serving as the update set) did not reveal clear patterns that differentiated these timeframes from those for which the testing procedure did not recommend updating. Nevertheless, calibration improved immediately after these updates.
Figure 2. Overall calibration by modeling approach and annual updating strategy. Left panels display calibration curves across the range of predictions produced by each model; right panels zoom in on calibration curves for predicted probabilities below 30%, which includes over 95% of all observations.
Figure 3. Calibration over time by modeling approach and updating strategy (smaller values are better). Dotted vertical lines highlight points at which the testing procedure recommended recalibration.

Discussion

We evaluated the impact of three competing updating strategies on performance of models for 30-day mortality after hospital admission over 7 years following initial model development. In addition to common strategies of retaining the original model or routinely refitting the model, we included a new data-driven strategy based on a nonparametric testing procedure for selecting among competing updating methods. This testing procedure is applicable regardless of the learning algorithm underlying the model, allowing our study to compare updating requirements of parallel LR, L1, RF, and NN models.

Updating requirements varied across modeling methods, both in terms of the timing and extent of updates. One year after model development, the nonparametric testing procedure recommended updating of all four models. These initial adjustments lead to immediate improvements in calibration in the months following update. Subsequent updating recommendations were varied and less frequent. The adjusted LR model was further recalibrated every other year with additional intercept corrections. The initial recalibration of the L1 model was retained until the intercept was further adjusted for the final two years of the study. The most significant and frequent updating was recommended for the NN model, which exhibited the least calibration drift over time. The testing procedure recommended refitting each year due to quite small improvements in the Brier score (~0.0001) compared to other updating approaches. As the Brier score takes into consideration both discrimination and calibration, the improvement in both dimensions of performance that resulted from refitting the NN model may have driven this recommendation. Refitting of the other models impacted calibration but did not significantly improve discrimination.

Some form of updating was warranted for all models. Retaining the original model over the course of the study period resulted in inferior calibration compared to routine refitting and test-based updating. Calibration measures of the original NN model did not exhibit significant trends indicative of calibration drift over the course of the study. Nevertheless, refitting this model each year, either as planned or as recommended by the testing procedure, still improved overall calibration, reduced month-to-month variability in calibration, and improved discrimination. Test-based updating of the other models improved upon the simple refitting strategy. Refitting corrected performance drift in the LR and L1 models; however, test-based updating recommendations resulted in lower ECIs (i.e., better calibration) and less month-to-month variability in performance compared to refitting. Refitting the RF model
improved overall calibration compared to the original model, but still resulted in variable calibration over shorter periods and did not correct performance drift over time. On the other hand, the test-based strategy avoided performance drift of the RF model and periods of instability observed under the refitting strategy (e.g., 2012) despite no additional updates being recommended.

In some cases, differences in calibration metrics between updating strategies were small and may not be clinically meaningful in practice. Whether these improvements are clinically meaningful in addition to being statistically significant is an important consideration and an open question for model comparison and impact assessment work. Although small in magnitude, the improvements in calibration under the test-based strategy compared to the refitting strategy highlight how recalibration may be sufficient, or even superior, to the standard practice of undertaking more substantial change by refitting. In addition, impact from recalibration is most likely to occur when patients are scored near user-defined cut-points that are clinically relevant, and assessment of clinically meaningful risk category reclassification anchors around what proportion of patients are near the cut-points (and change classification after calibration degredation).

These findings underscore a need for data-driven maintenance plans for clinical prediction models. A “one-size fits all” updating strategy will not suffice for all models. We cannot assume a new model built on recent data will be more generalizable to and perform better in the next cohort of patients than an existing model, even when large datasets, such as those in this study, are used for updating. Although calibration improved over the entire study period by regularly refitting the RF model, the model built on 2008 admissions did not improve upon, and may have actually performed worse, than the original RF model when applied in 2009. Similarly, we should not assume refitting is superior to simpler updating through recalibration. The intermittent recalibrations recommended by the testing procedure lead to better performance across the study period than routine refitting, both overall and on a month-to-month basis. Tailoring updating methods through data-driven updating strategies may therefore extend the accuracy and subsequent utility of prediction models beyond what might be achieved through simpler maintenance plans. We note, however, that these results may be sensitive to the volume of data available for updating, and further investigation regarding the impact of sample size is warranted.

Our results also highlight differences in the frequency with which models require updating. Despite being applied to the same data and therefore exposed to the same shifts in patient case mix and clinical environments, the LR, L1, RF, and NN models required updating at different time points. With the exception of the NN model, updating on an annual basis was not indicated and annual refits did not provide additional benefits over less frequent updates. Thus, we may experience inefficiencies under model maintenance plans requiring updates on pre-planned schedule. On the other hand, prescheduled updating plans may also neglect to update models in a timely manner, allowing periods of performance drift to go unnoticed and uncorrected. The cost of interim periods of reduced model accuracy may be difficult to assess as the prediction errors may impact patient outcomes, user confidence, and clinical efficiency. As health systems seek to implement clinical prediction more broadly and begin managing many prediction models, additional data-driven methods to determine when models require attention may be necessary and would complement maintenance strategies implementing test-based updating methods.

There are several limitations of the analyses presented here. We evaluated the three updating strategies in one clinical use case and population. Exploring how these updating strategies perform on models subject to different patterns of shifts in the clinical environment would provide more generalizable understanding. In this study, we limited the nonparametric testing procedure to consider five updating methods – retention of the existing model, intercept correction, linear logistic recalibration, flexible logistic recalibration, and model refitting. These updating methods are common and applicable across models; however, additional updating methods, some of which may be specific to certain learning algorithms, could easily be incorporated into the testing procedure. The availability of additional updating methods may impact when and how the test-based strategy adjusted the models over time. Further, we did not explore the impact of sample size. The volume of data available for constructing updates could have important impacts on both the model refitting and test-based updating strategies. For small samples, overfitting becomes more of a concern for the refitting strategy, while overly conservative updates may be a concern for the test-based strategy. We also acknowledge that the test-based updating strategy may be computationally intensive. Leveraging advances in computational resources and refining the number of bootstrap iterations considered in the first bootstrapping stage may reduce any computational burden. Tailoring the number of bootstrap iterations may also allow users to match statistical significance to clinically relevant magnitudes of change. Finally, all three of the updating strategies considered here may be inappropriate in the presence of significant changes in clinical practice or record systems that may render existing prediction models invalid. Any updating strategy must be flexible, both in terms of timing and approach, in response to such situations.

1008
Conclusion

We illustrated the use of a new data-driven updating strategy for clinical prediction models based on a variety of underlying modeling methods and compared this strategy to two baseline approaches in which models are either never updated or regularly refit on recent observations. The test-based updating strategy conservatively adjusted most models by recommending intermittent recalibration rather than repeated model refitting. Despite making limited adjustments to the models, the test-based updating strategy lead to more highly calibrated predictions than either of the baseline strategies. The test-based approach also highlighted differences in the updating requirements of common biostatistical and machine learning models, both in terms of the extent and timing of updates. These results have important implications for the implementation of clinical prediction models and the design of model maintenance plans. As the volume, complexity, and variability of prediction models implemented in health systems grows, data-driven updating policies could support model developers and managers as they endeavor to provide more stable and accurate model performance. In this way, data-driven updating strategies, such as the test-based approach presented here, will become key components of automated surveillance procedures that promote the long-term performance and utility of prediction models underlying a variety of informatics applications for decision support and population management.

Acknowledgements

Financial support for this study was provided by grants from the National Library of Medicine (5T15LM007450); the Veterans Health Administration (VA HSR&D IIR 13-052); and the National Institutes of Health (BCHI-R01-130828).

References

Mining Drugs and Indications for Suicide-Related Adverse Events

Tiffany Ding and Elizabeth S. Chen, PhD
Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

There has been a significant increase in suicide rates in the United States (U.S.) over the past two decades. Studies have highlighted the need for further exploration of suicide risk factors, particularly combinations of factors. In this study, a pharmacovigilance analysis was conducted to better understand drugs and indications as risk factors for suicide using data from the U.S. Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) and Adverse Event Open Learning through Universal Standardization (AEOLUS), a standardized version of FAERS. Association rule mining techniques were applied to 85,071 cases involving suicide-related adverse reactions and demographic subsets of these cases. Preliminary results reveal combinations of drugs and indications that may increase the likelihood of suicide, with certain combinations potentially affecting some demographic groups more than others. Further work is needed to validate the initial findings, explore subpopulations, and determine the broader implications for suicide prevention.

Introduction

Suicide is the 10th leading cause of death nationally.1 With suicide rates increasing nearly 30% from 1999 to 2016, there is a growing need for a comprehensive approach to suicide prevention.2 A key component of this approach is identifying risk factors to better understand and prevent suicidal thoughts and behaviors.3 One realm in which there is still much to be learned is which combinations of drugs or indications lead to an adverse drug reaction (ADR) of suicide and how these combinations may differ across various demographic groups.

With incidence rates of serious ADRs reaching 6.5%, ADRs have been identified as an important area of study.4 The FDA Adverse Event Reporting System (FAERS), a database containing post-marketing reactions for drugs in the United States, has facilitated the study of ADRs.5,6 There is much that can be discovered by mining this database for correlations between specific drugs and suicide-related outcomes. Previous studies using FAERS to investigate drugs and suicide have looked at the relationship between isotretinoin usage and depression and suicide, finasteride usage and suicidal behaviors, and suicidal ideation and suicidal behavior as ADRs for antidepressant drugs.7,9

One challenge in using FAERS is the existence of duplicate cases and misspellings, which makes it necessary to perform extensive data cleaning.10 In 2015, Banda et al. addressed this issue by creating a publicly available version of FAERS that removes duplicate reports, maps events to standardized concept identifiers (drug names to RxNorm concepts and outcomes to SNOMED-CT concepts), and precomputes summary statistics.11 This database, called the Adverse Event Open Learning through Universal Standardization (AEOLUS), was created using FAERS data from January 2004 to June 2015. A limitation of AEOLUS is that it does not preserve the demographic data present in FAERS, so it is necessary to use FAERS as a supplementary resource to retrieve patients’ demographic information.

Through basic analysis of adverse event databases, individual drugs that lead to a specific reaction can be discovered. To discover combinations of drugs that lead to a specific reaction, data mining techniques such as association rule mining (ARM), which generates association rules between items in a database, can be applied.12 An efficient method to find association rules in a database is via the Apriori algorithm.13 Applying ARM to adverse event reporting data can help uncover hidden drug-drug interactions that have serious effects on patients. Previous studies show that applying ARM to FAERS or other spontaneous reporting systems is a promising avenue for knowledge discovery.14 Yildirim applied ARM to FAERS to discover rules associating demographic information with adverse events for cases involving ciprofloxacin.15 Harpaz et al. performed ARM on FAERS data from 2008 to find multi-item associations between drugs and ADRs.16 Guo et al. used ARM to investigate the ADRs of two specific drugs, gadoversetamide and rofecoxib.17

The objective of this study is to apply ARM to cases in FAERS that have a suicide-related ADR in order to determine what combinations of drugs and indications are associated with suicide-related ADRs. Identifying these common combinations could provide guidance for further investigations of the effect of such combinations on suicide risk. Additionally, this preliminary study investigates how these association rules differ between male and female patients as well as across patients in different age groups, which can potentially contribute to informing personalized risk assessment.
Methods

Figure 1 depicts an overview of the study approach, which is based on the Knowledge Discovery in Databases process. The four major steps involved are: (1) Data selection to identify cases with a suicide-related reaction, (2) Data processing and subsetting to split the dataset into distinct demographic groups to allow for more detailed analyses, (3) Data mining using ARM, and (4) Interpretation and evaluation of results using visualization methods and comparison to relevant scientific literature. Data extraction, analysis, and visualization were conducted using the Julia general purpose programming language, MySQL database management system, and Tableau.

**Data selection**

The publicly available AEOLUS database, a standardized form of FAERS, was used as the primary data source. AEOLUS removes duplicate cases and applies standardized vocabulary to map FAERS labels to RxNorm Concept Unique Identifiers (for drugs) and SNOMED-CT identifiers (for indications). Additionally, AEOLUS precomputes some basic statistics such as the proportional reporting ratio (PRR). PRR is calculated as $\text{PRR} = \frac{a}{a+b} / \frac{c}{c+d}$, where $a$ is the number of cases with the suspected drug and suspected ADR, $b$ is the number of reports with the suspected drug and without the suspected ADR, $c$ is the number of reports without the suspected drug and with the suspected ADR, and $d$ is the number of reports without the suspected drug and without the suspected ADR.

The version of AEOLUS used for this study was created using FAERS data from January 2004 to June 2015. The format of the FAERS data collected from January 2004 to August 27, 2012 differs slightly from the format for data collected from September 2012 onwards in that the earlier data identify cases using isr, while the more recent data identify cases using primaryid. Consequently, the AEOLUS database keeps track of both isr and primaryid.

The AEOLUS concept table was queried for all concepts with names that match “%suicid%”. This query returned the following concepts: “depression suicidal,” “completed suicide,” “suicidal behavior,” “suicidal ideation,” “suicide attempt,” and “suicide of relative.” The first five concepts were identified as suicide-related ADRs and were chosen to be used to extract relevant cases for this study. “Suicide of relative” was left out because it is not an adverse reaction experienced by patients themselves.

The identifiers (primaryid for cases prior to August 2012; isr for cases after August 2012) of cases with reactions of “depression suicidal,” “completed suicide,” “suicidal behavior,” “suicidal ideation,” “suicide attempt,” and “suicide of relative” were extracted. These identifiers were used to obtain: (1) the list of drugs and (2) the list of indications for patients who experienced one or more of these reactions. The AEOLUS tables used to extract these data were standard_case_drug and standard_case_indication, respectively. The resulting dataset is referred to as the “full dataset.” Basic statistics regarding the drugs involved in these cases were extracted from the AEOLUS table standard_drug_outcome_statistics, which contains precomputed PRRs and frequencies for each drug-reaction combination.
Data processing and subsetting

The full dataset was segmented according to two demographic factors: (1) sex and (2) age. The purpose of data subsetting was to amplify signal detection of factors that increase suicide risk for specific population groups. AEOLUS does not contain demographic data, but by joining the AEOLUS and FAERS databases using primaryid and isr, demographic data were retrieved for the cases used in this study. Separate lists of drugs and indications were generated for female and male patients who had one of the previously mentioned suicide-related ADRs. Cases without a gender code were excluded. Similar lists were generated for patients in the following age groups: <15, 15-24, 25-34, 35-44, 45-54, 55-64, and >64. The selection of age groups follows the practices of the Centers for Disease Control and Prevention, with the exception that all age ranges less than 15 years old have been combined into one age group because children < 15 years old all tend to exhibit low rates of suicide. Cases with no age field and cases with ages measured in units other than years were excluded.

Data mining

A Julia package (ARules.jl) was used to perform ARM via the Apriori algorithm for the ten datasets (full dataset and nine subsets). The support threshold was set at 0.005 and the confidence threshold was set at 0.01. Support was calculated using the equation \( \text{supp}(X \rightarrow Y) = \frac{P(X \cup Y)}{n} \), confidence using the equation \( \text{conf}(X \rightarrow Y) = \frac{P(X | Y)}{P(X)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \), and lift using the equation \( \text{lift}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)\text{supp}(Y)} \). The maximum rule length was set to six to allow for the discovery of more complex rules should they exist.

Interpretation and evaluation

Tableau Desktop 2018.2 was used to visualize and organize the top drugs, top indications, and association rules discovered in this study. Association rules were grouped by RhS (right-hand side) to make it easier to see differences between the association rules for each demographic group.

PubMed and Side Effect Resource (SIDER), a centralized database of known ADRs that appear in drug labels, were used as resources to validate findings involving ADRs for individual drugs, SIDER 4.1, which was released in October 2015, was the version used in this study. PubMed was also used to validate findings involving indications or combinations of drugs.

Results

The characteristics of patients who had one or more suicide-related ADRs are described in Table 1. In total, AEOLUS contains 85,071 cases that involve a suicide-related ADR, with the most common suicide-related ADRs being completed suicide, suicidal ideation, and suicide attempt. Note that some cases are associated with more than one suicide-related ADR, so the percentages for each ADR add up to more than 100%.

Table 1. Description of patient cohort used in analysis of suicide-related ADRs.

<table>
<thead>
<tr>
<th>ADR</th>
<th># cases</th>
<th>% cases</th>
<th>Age group</th>
<th># cases</th>
<th>% cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed suicide</td>
<td>32,337</td>
<td>38.01%</td>
<td>&lt; 15</td>
<td>7,396</td>
<td>8.69%</td>
</tr>
<tr>
<td>Depression suicidal</td>
<td>1,062</td>
<td>1.25%</td>
<td>15 - 24</td>
<td>8,425</td>
<td>9.90%</td>
</tr>
<tr>
<td>Suicidal behavior</td>
<td>1,494</td>
<td>1.76%</td>
<td>25 - 34</td>
<td>9,774</td>
<td>11.49%</td>
</tr>
<tr>
<td>Suicidal ideation</td>
<td>32,227</td>
<td>37.88%</td>
<td>35 - 44</td>
<td>12,783</td>
<td>15.03%</td>
</tr>
<tr>
<td>Suicide attempt</td>
<td>21,298</td>
<td>25.03%</td>
<td>45 - 54</td>
<td>14,153</td>
<td>16.63%</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>31,156</td>
<td>36.62%</td>
<td>&gt; 65</td>
<td>5,768</td>
<td>6.78%</td>
</tr>
<tr>
<td>Female</td>
<td>39,484</td>
<td>46.41%</td>
<td>Non-year unit</td>
<td>635</td>
<td>0.75%</td>
</tr>
<tr>
<td>Unknown gender</td>
<td>17,778</td>
<td>20.90%</td>
<td>Unknown age</td>
<td>17,236</td>
<td>20.26%</td>
</tr>
</tbody>
</table>

Table 2 shows the top drugs associated with cases involving suicide-related ADRs. The left-hand side of the table shows the top ten drugs ranked by PRR using a cutoff of count ≥ 3 to highlight stronger signals, while the right-hand side of the table shows the top ten drugs ranked by case count. The “S” superscript following a drug name denotes that the drug has a suicide-related side effect (suicide, suicidal ideation, suicide attempt, completed suicide, or suicidal behavior) in the SIDER database. The “P” superscript denotes that the drug does not have a documented suicide-related side effect in the SIDER database, but there exists at least one PubMed article linking the drug to a suicide-related reaction.
Table 2. Top drugs associated with cases with suicide-related ADRs ranked by PRR (left) and count (right).

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Reaction</th>
<th>PRR</th>
<th>Count</th>
<th>Drug Name</th>
<th>Reaction</th>
<th>PRR</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon Black</td>
<td>Completed suicide</td>
<td>244.91289</td>
<td>12</td>
<td>Varenicline</td>
<td>Suicidal ideation</td>
<td>12.08389</td>
<td>4069</td>
</tr>
<tr>
<td>Carbaryl</td>
<td>Completed suicide</td>
<td>147.95526</td>
<td>3</td>
<td>Acetaminophen</td>
<td>Completed suicide</td>
<td>3.58407</td>
<td>4025</td>
</tr>
<tr>
<td>Atropine / Hyoscyamine / Phenobarbital / Scopolamine Oral Solution</td>
<td>Completed suicide</td>
<td>134.50738</td>
<td>5</td>
<td>Alprazolam</td>
<td>Completed suicide</td>
<td>5.54899</td>
<td>3504</td>
</tr>
<tr>
<td>Acetaminophen / Dextromethorphan / Doxylamine Oral Tablet</td>
<td>Completed suicide</td>
<td>113.334</td>
<td>9</td>
<td>Paroxetine</td>
<td>Suicidal ideation</td>
<td>6.09995</td>
<td>2780</td>
</tr>
<tr>
<td>Captan</td>
<td>Completed suicide</td>
<td>112.7289</td>
<td>4</td>
<td>Quetiapine</td>
<td>Completed suicide</td>
<td>5.25461</td>
<td>2621</td>
</tr>
<tr>
<td>Ethylene Glycol</td>
<td>Completed suicide</td>
<td>112.43552</td>
<td>144</td>
<td>EthanolP, 28</td>
<td>Completed suicide</td>
<td>58.77839</td>
<td>2482</td>
</tr>
<tr>
<td>Acetaminophen / Caffeine / Pyrilamine Oral Capsule</td>
<td>Completed suicide</td>
<td>110.96644</td>
<td>3</td>
<td>Acetaminophen</td>
<td>Hydrocodone Oral Tablet</td>
<td>Completed suicide</td>
<td>24.60013</td>
</tr>
<tr>
<td>Helium</td>
<td>Completed suicide</td>
<td>105.68436</td>
<td>5</td>
<td>Varenicline</td>
<td>Suicide attempt</td>
<td>9.58679</td>
<td>2216</td>
</tr>
<tr>
<td>Konjac Mannan</td>
<td>Completed suicide</td>
<td>103.57267</td>
<td>7</td>
<td>Zolpidem</td>
<td>Completed suicide</td>
<td>3.83457</td>
<td>2169</td>
</tr>
<tr>
<td>Carbon Monoxide</td>
<td>Completed suicide</td>
<td>102.87414</td>
<td>112</td>
<td>Clonazepam</td>
<td>Completed suicide</td>
<td>3.94694</td>
<td>2144</td>
</tr>
</tbody>
</table>

found in SIDER; P, # found in PubMed (where # is the reference to the PubMed article)

Of the 85,071 cases with suicide-related ADRs, 44,782 cases (52.64%) have one or more known indications. 15,617 cases (18.36%) were tagged with only “product used for unknown indication” or “drug use for unknown indication,” while the remaining cases had no indication information. Figure 2 depicts the indications that appear most frequently in the 44,782 cases with known indications. The ten most frequent indications are depicted in their own circles, with the remaining indications grouped together in the circle marked “other.”

Figure 2. Top indications associated with cases with suicide-related ADRs.

Table 3. Top association rules for drugs in all cases with suicide-related ADRs, grouped by Lhs and ordered by descending lift.

<table>
<thead>
<tr>
<th>Rhs</th>
<th>Lhs</th>
<th>Conf</th>
<th>Lift</th>
<th>F</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen</td>
<td>{Carboxylic acid}</td>
<td>0.26</td>
<td>3.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Varenicline</td>
<td>{Hydrogenated}</td>
<td>0.22</td>
<td>3.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clonazepam</td>
<td>{Doxylamine}</td>
<td>0.19</td>
<td>2.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribavirin</td>
<td>{Riboside}</td>
<td>0.05</td>
<td>2.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peginterferon alfa-2a</td>
<td>{Hyaluronate}</td>
<td>0.25</td>
<td>3.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peginterferon alfa-2a</td>
<td>{Riboside}</td>
<td>0.19</td>
<td>2.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peginterferon alfa-2a</td>
<td>{Ribose}</td>
<td>0.19</td>
<td>2.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peginterferon alfa-2a</td>
<td>{Ribonucleic acid}</td>
<td>0.19</td>
<td>2.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peginterferon alfa-2a</td>
<td>{Lamotrigine}</td>
<td>0.19</td>
<td>2.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peginterferon alfa-2a</td>
<td>{Riboside}</td>
<td>0.18</td>
<td>2.48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Acetaminophen/Hydrocodone Oral Tablet
Of the 85,071 cases with suicide-related ADRs, 84,330 (99.13%) contained drug data. Table 3 displays the association rules for these cases. Rules should be interpreted as Lhs (left-hand side) → Rhs (right-hand side). For example, the first row of Table 3 should be read as {Ibuprofen} → Acetaminophen, meaning that among patients who had suicide-related ADRs, those who were taking ibuprofen were likely to be also taking acetaminophen. The left-hand side of the rule can contain more than one element, but the top drug association rules that were discovered all have left-hand sides that are one-element sets. In Tables 3-9, only rules with conf ≥ 0.18 and lift ≥ 2 are displayed. The results are grouped by Rhs and ranked from greatest to least lift value within each Rhs. Support and confidence tend to exaggerate the importance of rules involving items that appear frequently in the database even if the relationship between items is weak, but lift attempts to correct for frequency and is thus chosen as the primary measure of rule interestingness in this study.29

The association rules for indications in the 44,782 cases with known indications with conf ≥ 0.18 and lift ≥ 2 are as follows: {depression | smoking cessation therapy} → anxiety (conf = 0.458, lift = 6.083), {sleep disorder} → anxiety (conf = 0.279, lift = 3.707), {insomnia} → anxiety (conf = 0.242, lift = 3.209), {anxiety | smoking cessation therapy} → depression (conf = 0.597, lift = 2.698), {anxiety} → depression (conf = 0.490, lift = 2.212). Unlike in Table 3, some of these rules contain left-hand sides that involve more than one element. For example, the first rule implies that among patients who experienced a suicide-related ADR, those who have indications of both depression and smoking cessation therapy are also likely to have an indication of anxiety.

Tables 4 and 5 show the differences between drug association rules for the 39,095 suicide-related ADR cases that involve female patients and contain drug data vs. the 30,035 cases involving male patients. Tables 6 and 7 show indication association rules for the 21,533 suicide-related ADR cases that involve female patients and contain indications data vs. the 17,721 cases involving male patients.

**Table 4.** Top association rules for drugs in cases involving female patients with suicide-related ADRs, grouped by Rhs and ordered by descending lift.

<table>
<thead>
<tr>
<th>Rhs</th>
<th>Lhs</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen</td>
<td>{Ibuprofen}</td>
<td>0.268</td>
<td>3.302</td>
</tr>
<tr>
<td></td>
<td>{Diphenhydramine}</td>
<td>0.238</td>
<td>2.938</td>
</tr>
<tr>
<td></td>
<td>{Oxycodeine}</td>
<td>0.199</td>
<td>2.451</td>
</tr>
<tr>
<td></td>
<td>{Ethanol}</td>
<td>0.188</td>
<td>2.315</td>
</tr>
<tr>
<td>Alprazolam</td>
<td>{Carisoprodil}</td>
<td>0.273</td>
<td>3.282</td>
</tr>
<tr>
<td></td>
<td>{Acetaminophen*}</td>
<td>0.220</td>
<td>2.636</td>
</tr>
<tr>
<td></td>
<td>{Hydrocodone}</td>
<td>0.204</td>
<td>2.452</td>
</tr>
<tr>
<td></td>
<td>{Zolpidem}</td>
<td>0.192</td>
<td>2.300</td>
</tr>
<tr>
<td></td>
<td>{Oxycodeine}</td>
<td>0.185</td>
<td>2.218</td>
</tr>
<tr>
<td>Clonazepam</td>
<td>{Trazodone}</td>
<td>0.219</td>
<td>3.185</td>
</tr>
<tr>
<td></td>
<td>{Valproate}</td>
<td>0.262</td>
<td>3.078</td>
</tr>
<tr>
<td></td>
<td>{Risperidone}</td>
<td>0.230</td>
<td>2.706</td>
</tr>
<tr>
<td></td>
<td>{Lamotrigine}</td>
<td>0.214</td>
<td>2.512</td>
</tr>
<tr>
<td></td>
<td>{Clonazepam}</td>
<td>0.210</td>
<td>2.468</td>
</tr>
<tr>
<td></td>
<td>{Lorazepam}</td>
<td>0.204</td>
<td>2.398</td>
</tr>
<tr>
<td></td>
<td>{Trazodone}</td>
<td>0.187</td>
<td>2.201</td>
</tr>
</tbody>
</table>

*Acetaminophen/Hydrocodone Oral Tablet

**Table 5.** Top association rules for drugs in cases involving male patients with suicide-related ADRs, grouped by Rhs and ordered by descending lift.

<table>
<thead>
<tr>
<th>Rhs</th>
<th>Lhs</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen</td>
<td>{Ibuprofen}</td>
<td>0.23</td>
<td>4.08</td>
</tr>
<tr>
<td></td>
<td>{Alprazolam}</td>
<td>0.24</td>
<td>3.71</td>
</tr>
<tr>
<td></td>
<td>{Hydrocodone}</td>
<td>0.19</td>
<td>3.95</td>
</tr>
<tr>
<td></td>
<td>{Oxycodeine}</td>
<td>0.19</td>
<td>6.79</td>
</tr>
<tr>
<td></td>
<td>{Metoprolol}</td>
<td>0.22</td>
<td>8.02</td>
</tr>
<tr>
<td></td>
<td>{Lisinopril}</td>
<td>0.19</td>
<td>6.06</td>
</tr>
<tr>
<td></td>
<td>{Aspirin}</td>
<td>0.26</td>
<td>4.37</td>
</tr>
<tr>
<td></td>
<td>{Laronidazol}</td>
<td>0.22</td>
<td>3.47</td>
</tr>
<tr>
<td></td>
<td>{Lisinopril}</td>
<td>0.20</td>
<td>6.06</td>
</tr>
<tr>
<td></td>
<td>{Ribavirin}</td>
<td>0.48</td>
<td>65.42</td>
</tr>
<tr>
<td></td>
<td>{Ribavirin}</td>
<td>0.38</td>
<td>66.36</td>
</tr>
<tr>
<td></td>
<td>{Valproate}</td>
<td>0.26</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>{Trazodone}</td>
<td>0.23</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td>{Lamotrigine}</td>
<td>0.21</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td>{Olanzapine}</td>
<td>0.19</td>
<td>2.64</td>
</tr>
<tr>
<td></td>
<td>{Clonazepam}</td>
<td>0.18</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td>{Ribavirin}</td>
<td>0.91</td>
<td>65.42</td>
</tr>
<tr>
<td></td>
<td>{Ribavirin}</td>
<td>0.91</td>
<td>65.36</td>
</tr>
<tr>
<td></td>
<td>{Aspirin}</td>
<td>0.19</td>
<td>8.02</td>
</tr>
</tbody>
</table>
Table 6. Top association rules for indications in cases involving female patients with suicide-related ADRs, grouped by Rhs and ordered by descending lift.

<table>
<thead>
<tr>
<th>Rhs</th>
<th>Lhs</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>anxiety</td>
<td>[depression</td>
<td>smoking cessation therapy]</td>
<td>0.466</td>
</tr>
<tr>
<td></td>
<td>[pain</td>
<td>smoking cessation therapy]</td>
<td>0.420</td>
</tr>
<tr>
<td></td>
<td>[depression</td>
<td>insomnia]</td>
<td>0.387</td>
</tr>
<tr>
<td></td>
<td>[depression</td>
<td>pain]</td>
<td>0.381</td>
</tr>
<tr>
<td></td>
<td>[bipolar disorder</td>
<td>depression]</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>[sleep disorder]</td>
<td>0.303</td>
<td>3.678</td>
</tr>
<tr>
<td></td>
<td>[insomnia]</td>
<td>0.272</td>
<td>3.297</td>
</tr>
<tr>
<td></td>
<td>[pain]</td>
<td>0.181</td>
<td>2.195</td>
</tr>
</tbody>
</table>

Table 7. Top association rules for indications in cases involving male patients with suicide-related ADRs, grouped by Rhs and ordered by descending lift.

<table>
<thead>
<tr>
<th>Rhs</th>
<th>Lhs</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>anxiety</td>
<td>[depression</td>
<td>smoking cessation therapy]</td>
<td>0.453</td>
</tr>
<tr>
<td></td>
<td>[insomnia]</td>
<td>0.210</td>
<td>3.043</td>
</tr>
<tr>
<td>depression</td>
<td>[anxiety]</td>
<td>0.586</td>
<td>3.006</td>
</tr>
<tr>
<td></td>
<td>[anxiety]</td>
<td>0.456</td>
<td>2.941</td>
</tr>
<tr>
<td>smoking cessation therapy</td>
<td>[gastroesophageal reflux disease]</td>
<td>0.522</td>
<td>4.184</td>
</tr>
<tr>
<td></td>
<td>[hypertension]</td>
<td>0.264</td>
<td>2.115</td>
</tr>
</tbody>
</table>

Association rules for drugs were found to differ across age groups. Table 8 displays association rules for the 8,317 cases that have drug data and involve a patient between 15 and 24 years old. Table 9 displays the association rules for the 14,153 cases that have drug data and involve a patient between 45 and 54 years old. These two age groups are particularly worthy of investigation because suicide is a significant cause of death in both of these age groups. In 2016, suicide was the second leading cause of death among 15-24 year old Americans (436 cases) and the fourth leading cause of death among 45-55 year old Americans (8,437 cases).

The only association rule with conf ≥ 0.18 and lift ≥ 2 that resulted from mining the 4,463 cases that have indications data and involve a patient between 15 and 24 years old is \{insomnia\} → anxiety (conf = 0.280, lift = 4.287). Table 10 displays association rules for the 7,138 cases that have indications data and involve a patient between 45 and 54 years old. Many significant association rules were discovered in this age group, so in order to highlight the most interesting rules, only rules with conf ≥ 0.25 and lift ≥ 4 are displayed.

Table 8. Top association rules for drugs in cases involving patients 15-24 years old who experienced suicide-related ADRs, grouped by Rhs and ordered by descending lift.

<table>
<thead>
<tr>
<th>Rhs</th>
<th>Lhs</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen</td>
<td>{Aspirin}</td>
<td>0.366</td>
<td>4.782</td>
</tr>
<tr>
<td></td>
<td>{Ibuprofen}</td>
<td>0.310</td>
<td>4.047</td>
</tr>
<tr>
<td>Alprazolm</td>
<td>{Carisoprodol}</td>
<td>0.557</td>
<td>9.449</td>
</tr>
<tr>
<td></td>
<td>{Oxycodone}</td>
<td>0.352</td>
<td>5.982</td>
</tr>
<tr>
<td></td>
<td>{Methadone}</td>
<td>0.276</td>
<td>4.690</td>
</tr>
<tr>
<td></td>
<td>{Acetaminophen}*</td>
<td>0.744</td>
<td>4.145</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>{Aspirin}</td>
<td>0.308</td>
<td>5.798</td>
</tr>
<tr>
<td></td>
<td>{Acetaminophen}</td>
<td>0.215</td>
<td>4.047</td>
</tr>
<tr>
<td>Quetiapine</td>
<td>{Valproate}</td>
<td>0.281</td>
<td>4.236</td>
</tr>
<tr>
<td></td>
<td>{Lamotrigine}</td>
<td>0.244</td>
<td>3.669</td>
</tr>
<tr>
<td></td>
<td>{Olanzapine}</td>
<td>0.196</td>
<td>2.959</td>
</tr>
<tr>
<td></td>
<td>{Lorazepam}</td>
<td>0.189</td>
<td>2.851</td>
</tr>
</tbody>
</table>

*Acetaminophen/Hydrocodone Oral Tablet
Table 9. Top association rules for drugs in cases involving 45-54 years old who experienced suicide-related ADRs, grouped by Rhs and ordered by descending lift.

<table>
<thead>
<tr>
<th>Rhs</th>
<th>Lhs</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetaminophen</td>
<td>{Ibuprofen}</td>
<td>0.25</td>
<td>3.18</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>{Aspirin}</td>
<td>0.19</td>
<td>2.46</td>
</tr>
<tr>
<td>Acetaminophen*</td>
<td>{Carisoprodol}</td>
<td>0.22</td>
<td>5.78</td>
</tr>
<tr>
<td>Alprazolam</td>
<td>{Carisoprodol}</td>
<td>0.28</td>
<td>2.87</td>
</tr>
<tr>
<td></td>
<td>{Hydrocodone}</td>
<td>0.25</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>{Temazepam}</td>
<td>0.23</td>
<td>2.40</td>
</tr>
<tr>
<td></td>
<td>{Methadone}</td>
<td>0.22</td>
<td>2.81</td>
</tr>
<tr>
<td></td>
<td>{Oxycodone}</td>
<td>0.22</td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td>{clobenzapine}</td>
<td>0.20</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td>{zolpidem}</td>
<td>0.20</td>
<td>2.03</td>
</tr>
<tr>
<td>Amlodipine</td>
<td>{Lisinopril}</td>
<td>0.21</td>
<td>5.81</td>
</tr>
<tr>
<td>Bupropion</td>
<td>{Diphenhydramine}</td>
<td>0.22</td>
<td>3.14</td>
</tr>
<tr>
<td>Carisoprodol</td>
<td>{Acetaminophen}</td>
<td>0.18</td>
<td>5.78</td>
</tr>
<tr>
<td>Clonazepam</td>
<td>{lamotrigine}</td>
<td>0.25</td>
<td>3.34</td>
</tr>
<tr>
<td></td>
<td>{Trazodone}</td>
<td>0.19</td>
<td>2.68</td>
</tr>
<tr>
<td>Diazepam</td>
<td>{Temazepam}</td>
<td>0.23</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>{Morphine}</td>
<td>0.20</td>
<td>3.36</td>
</tr>
<tr>
<td>gabapentin</td>
<td>{Hydroxyzine}</td>
<td>0.25</td>
<td>3.41</td>
</tr>
<tr>
<td>Hydrocodone</td>
<td>{Carisoprodol}</td>
<td>0.21</td>
<td>4.24</td>
</tr>
<tr>
<td>Lisinopril</td>
<td>{Hydrochlorothiazide}</td>
<td>0.32</td>
<td>7.92</td>
</tr>
<tr>
<td></td>
<td>{Amlodipine}</td>
<td>0.23</td>
<td>5.81</td>
</tr>
<tr>
<td>peginterferon alfa-2a</td>
<td>{Ribavirin}</td>
<td>0.49</td>
<td>44.11</td>
</tr>
<tr>
<td>peginterferon alfa-2b</td>
<td>{Ribavirin}</td>
<td>0.38</td>
<td>45.43</td>
</tr>
<tr>
<td>quetiapine</td>
<td>{Lorazepam}</td>
<td>0.19</td>
<td>2.55</td>
</tr>
<tr>
<td></td>
<td>{Valproate}</td>
<td>0.19</td>
<td>2.53</td>
</tr>
<tr>
<td>Ribavirin</td>
<td>{peginterferon alfa-2a}</td>
<td>0.96</td>
<td>45.43</td>
</tr>
<tr>
<td></td>
<td>{peginterferon alfa-2b}</td>
<td>0.93</td>
<td>44.11</td>
</tr>
<tr>
<td>varenicline</td>
<td>{Albuterol}</td>
<td>0.40</td>
<td>3.63</td>
</tr>
<tr>
<td></td>
<td>{Simvastatin}</td>
<td>0.23</td>
<td>2.10</td>
</tr>
<tr>
<td>zolpidem</td>
<td>{Lorazepam}</td>
<td>0.18</td>
<td>2.97</td>
</tr>
</tbody>
</table>

*Acetaminophen/Hydrocodone Oral Tablet

Table 10. Top association rules for indications in cases involving patients 45-54 years old who experienced suicide-related ADRs, grouped by Rhs and ordered by descending lift.

<table>
<thead>
<tr>
<th>Rhs</th>
<th>Lhs</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>anxiety</td>
<td>{depression</td>
<td>pain</td>
<td>smoking cessation therapy}</td>
</tr>
<tr>
<td></td>
<td>{depression</td>
<td>smoking cessation therapy}</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>{hypertension</td>
<td>pain}</td>
<td>0.474</td>
</tr>
<tr>
<td></td>
<td>{depression</td>
<td>hypertension}</td>
<td>0.467</td>
</tr>
<tr>
<td></td>
<td>{depression</td>
<td>pain}</td>
<td>0.462</td>
</tr>
<tr>
<td></td>
<td>{depression</td>
<td>sleep disorder}</td>
<td>0.448</td>
</tr>
<tr>
<td></td>
<td>{bipolar disorder</td>
<td>depression}</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td>{pain</td>
<td>smoking cessation therapy}</td>
<td>0.427</td>
</tr>
<tr>
<td></td>
<td>{depression</td>
<td>insomnia}</td>
<td>0.425</td>
</tr>
<tr>
<td>hypertension</td>
<td>{blood cholesterol increased}</td>
<td>0.387</td>
<td>7.819</td>
</tr>
<tr>
<td></td>
<td>{gastroesophageal reflux disease}</td>
<td>0.365</td>
<td>7.398</td>
</tr>
<tr>
<td></td>
<td>{diabetes mellitus}</td>
<td>0.340</td>
<td>6.878</td>
</tr>
<tr>
<td></td>
<td>{anxiety</td>
<td>pain}</td>
<td>0.303</td>
</tr>
<tr>
<td>pain</td>
<td>{gastroesophageal reflux disease</td>
<td>smoking cessation therapy}</td>
<td>0.414</td>
</tr>
<tr>
<td></td>
<td>{anxiety</td>
<td>hypertension}</td>
<td>0.385</td>
</tr>
<tr>
<td></td>
<td>{anxiety</td>
<td>depression</td>
<td>smoking cessation therapy}</td>
</tr>
<tr>
<td></td>
<td>{gastroesophageal reflux disease}</td>
<td>0.359</td>
<td>4.495</td>
</tr>
<tr>
<td></td>
<td>{anxiety</td>
<td>smoking cessation therapy}</td>
<td>0.356</td>
</tr>
</tbody>
</table>

Discussion

In this preliminary study, FAERS cases involving suicide-related ADRs were used to generate basic statistics and association rules. ARM was performed on all cases with suicide-related ADRs, cases with male patients, cases with female patients, and cases split into seven different age groups.

Table 3 contains the association rules discovered by performing ARM on all cases with suicide-related ADRs. These rules are good starting points for future studies intending to investigate combinations of drugs that may lead to suicide. One example is the rule {Valproate} → quetiapine (conf = 0.26, lift = 3.36). Valproate and quetiapine have been found to increase the rate of self-harm and suicide events when taken individually, but there is yet to be a study of how suicide rates are affected by taking these two drugs concurrently.30

Tables 4 and 5 illustrate how data segmentation allows for the discovery of rules that are hidden when analyzing all patients as one group. For example, one association rule that is seen in the male population is {Amlodipine} → lisinopril (conf = 0.20, lift = 6.06). Amlodipine and lisinopril are often used in combination to treat hypertension, but the preliminary findings of this study suggest that other drugs could be considered when treating hypertension in male patients in order to mitigate the risk of suicide-related ADRs.31
Additionally, this study highlights sex differences when it comes to indications in cases involving suicide-related ADRs. Tables 6 and 7 show that “pain” appears frequently among the top association rules for indications for female patients (present in six out of 18 rules), but “pain” does not appear at all in the top association rules for indications for male patients. This is potentially due to greater pain sensitivity in females, but these findings nonetheless suggest that females experiencing pain concurrently with other indications present in the association rules in Table 6 could have an elevated risk of suicide-related ADRs.32

Drug association rules also differ across age groups. Table 8 shows that the top rules for drugs in cases involving 15 to 24-year-old patients are very similar to the top rules mined using all patients (Table 3), but Table 9 reveals association rules present in 45 to 54-year-old patients that were not significant in the general population. Examples of such rules include {zolpidem} → alprazolam (conf = 0.20, lift = 2.03), {temazepam} → diazepam (conf = 0.23, lift = 3.83), and {albuterol} → varenicline (conf = 0.40, lift = 3.63). Table 10 reveals groups of indications that may increase the likelihood of suicide-related ADRs for patients that are 45 to 55 years old. This study highlights that certain drugs and indications may have differing effects on the suicide risk of patients in different age groups.

One limitation of this study is that it is difficult to determine cause and effect relationships. In the FAERS and AEOLUS databases, it is difficult to determine whether a drug that is associated with a suicide-related ADR is a drug leading to suicidal behaviors or a drug purposefully taken to perform suicide. For example, acetaminophen appears in several association rules discovered in this study. However, since acetaminophen, more commonly known by its brand name Tylenol, is one of the most commonly abused drugs, it is possible that taking acetaminophen is a manifestation of suicidal behaviors rather than the cause.33 This suggests that some of the drugs that appear in the association rules mined in this study are drugs used in suicide attempts rather than drugs leading to suicide-related behavior. Future studies investigating suicide could supplement data from spontaneous reporting systems like FAERS with longitudinal data from electronic health records and claims data, which would provide better evidence for cause and effect relationships.

Another challenge of this study was determining appropriate minimum support and confidence levels for (1) running the Apriori algorithm and (2) determining which rules outputted by the Apriori algorithm are worthy of further investigation. In this study, metrics such as PRR, support, confidence, and lift were used for ranking individual drugs and associations. Additional metrics such as chi-square could be calculated and compared to assess performance. Since support and confidence tend to overlook rare rules and are thus imperfect measures of rule interestingness, implementing the Apriori algorithm using a different interestingness metric as a threshold is worthy of future exploration.34 Sindhu and Kannan have used the Apriori algorithm modified to utilize PRR instead of the traditional confidence as a threshold to mine FAERS.35 Ibrahim et al. have used a hybrid Apriori algorithm that uses PRR and chi-square as interestingness thresholds to mine FAERS.36 Applying these algorithms to data from different demographic segments, as was done in this study, would be a good next step.

Other next steps include studying how drugs and indications differ among various suicide-related ADRs and conducting formal evaluations. Early findings suggest that certain drugs are more likely to lead to some suicide-related ADRs than other suicide-related ADRs. For example, varenicline is associated with 4,069 cases of suicidal ideation in AEOLUS but only 681 cases of completed suicide, whereas acetaminophen is associated with only 876 cases of suicidal ideation but 4,025 cases of completed suicide. Formal evaluations will involve validating the results with clinical experts and established medical knowledge resources as well as determining if there is a statistically significant difference among demographic groups. For example, quantitative evaluations could involve comparisons of rules to characterize similarities as well as differences (e.g., using F-score or Matthews correlation coefficient). Additionally, in order to determine which combinations of drugs and indications contribute uniquely to suicide-related ADRs, a control group could be used (e.g., by removing association rules that appear in populations without suicide-related ADRs from those generated in this study). Potential applications of these findings for supporting clinical decisions and precision medicine will also be further explored (e.g., using longitudinal electronic health data to examine suicide risk for patients who are prescribed a given combination of drugs).37

Conclusion

There is a need to better understand risk factors for suicide. Using data from a spontaneous reporting system and data mining techniques, this study characterized combinations of drugs and indications associated with suicide-related ADRs. These initial findings can serve as a resource for informing investigations of potential drug-drug interactions or comorbidities for suicide risk. Further work is needed to validate the initial findings, explore subpopulations, and determine the broader implications for suicide prevention.
Acknowledgments

This work was funded in part by National Institutes of Health grant U54GM115677. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. WISQARS leading causes of death reports - CDC. Available at: https://webappa.cdc.gov/sasweb/ncipc/leadcause.html
23. Suicide Statistics – AFSP. Available at: https://afsp.org/about-suicide/suicide-statistics/
24. Julia Association Rules Package. Available at: https://github.com/bcbi/ARules.jl
An evaluation of mHealth adoption and health self-management in emerging adulthood

Connor Drake, MPA, PhD Candidate¹, Meagan Cannady MSPH², Kathryn Howley², Christopher Shea, PhD, MA, MPA³, Ralph Snyderman MD²

¹Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; Duke Center for Personalized Health Care, School of Medicine, Duke University, Durham, NC, USA; ²Duke Center for Personalized Health Care, School of Medicine, Duke University, Durham, NC, USA; ³Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

This study offers a description of factors that predict the adoption of mobile health technologies (mHealth) and their application for health self-management in emerging adults. Primary data collection occurred at three diverse post-secondary educational institutions (N= 1,329). The analysis used a logistic regression to identify predictors of mHealth adoption. Descriptive analyses are presented on health self-management applications and perceived ease of use and effectiveness. Use of mHealth was high in respondents (58.5%). Factors associated with increased likelihood of mHealth adoption included being female, overweight or obese, having a chronic condition, eating the recommended amount of daily fruit, and engaging in regular moderate exercise. Low household income was associated with being less likely to use mHealth. The most common self-management application for mHealth was for tracking physical activity. Findings related to ease of use and effectiveness ratings by applications may provide insight into designing more effective mHealth tools in this population.

Introduction

The development of digital technologies is spurring the expansion of mHealth, broadly defined as “health care and public health practices supported by mobile devices”¹. mHealth can be thought of as a subcategory of telehealth and includes health apps or mobile phone capabilities that enable immediate access to biometric data, health information, patient education, adherence reminders, or progress on a lifestyle modification goal.² The prevalence of consumer mHealth apps has grown dramatically. In 2017 there were over 318,000 health apps available for download, doubling the number available in 2015.³ This growth is due, in part, to the increasing prevalence of smart phones and watches. As of 2018, the vast majority of individuals in the United States have a cell phone (95%) and smartphone ownership grew dramatically from 35% in 2011 to 77% in 2018.⁴

There is evidence of promising potential for mHealth to improve health behaviors and help individuals self-manage chronic conditions. A systematic review of such interventions based primarily on randomized controlled trials found that mHealth intervention participants were more successful in changing a variety of health behaviors and behavior-related outcomes, including physical activity, smoking cessation, consuming a healthy diet, weight loss, medication adherence, improved blood pressure control, and better glycemic control.¹ Despite the potential benefits of mHealth, there are also significant barriers to widespread dissemination. Recent evidence suggests that as many as 45% of participants in a survey study reported that they downloaded health apps they no longer use.⁵ Furthermore, a 2017 study that explored sustained use for activity trackers found that, on average, participants used trackers for only 129 days.⁶ A separate study estimated that as many as one third of activity trackers were discontinued within 6 months of use.⁷ Successful adoption of mHealth will require a better understanding of the preferences of users and an emphasis on design to encourage sustained and effective use.⁸
With the increasing ubiquity of mHealth and health information technology, it is important to understand its use and impact on health, particularly for younger generations. Emerging adulthood is a distinct development period that occurs between 18 years and 25-30 years marked by an increase in autonomy of decision making and identity formation. Emerging adulthood is often associated with the height of health and well-being; however, recent trends indicate that the rate at which younger adults in the United States are being diagnosed with chronic medical and mental health conditions has grown dramatically. As a result, the periods of young and emerging adulthood provide opportunities for emphasizing primary prevention efforts to improve long term health outcomes. Enduring patterns of health behaviors and lifestyle choices are established during emerging adulthood, providing evidence that this period is uniquely relevant to health promotion and disease prevention efforts.

mHealth technologies, whether consumer chosen or health provider directed, will likely play an increasingly important role in health care. To this end, health care improvement efforts have focused on implementing approaches to care that leverage mHealth to support health self-management. Health self-management consists of the broad set of skills, knowledge, and abilities to manage one’s health and health care decision making across the spectrum from health enhancement or prevention to disease management. As a cost-effective strategy to improve health self-management, understanding predictors of mHealth technology use and factors associated with their effectiveness is becoming increasingly important. To this end, the technology acceptance model sheds light on how users accept and use a technology like mHealth. In addition to important contextual factors, perceived ease of use and effectiveness of a technology influences how and whether the user will adopt a specific technology.

Several distinct characteristics have been noted among users of mHealth. Users are younger, wealthier and work full time. The increased use of health apps was associated with the prevalence of chronic conditions, including obesity. Fitness and nutrition have been identified as the most commonly reported categories of health apps used. Of respondents that use a fitness or nutrition app, most use them at least daily. These findings across the literature were based on samples of adults and did not include a sub-group analysis on younger emerging adults.

The goal of this study was to define factors associated with mHealth use, to describe the intended health self-management applications, and determine their ease of use and effectiveness among emerging adults attending three different institutions of post-secondary education in the southeastern United States. The three primary research questions in this study were: 1) What contextual factors, behaviors, and characteristics of emerging adults are predictive of the likelihood of using mHealth? 2) For emerging adults that use mHealth, what health self-management activities are they used for? 3) What is the perceived ease of use and effectiveness of the mHealth for each health self-management activity?

There are notable gaps in the literature surrounding mHealth use in emerging adulthood. To address them, this study is the first to examine predictors of mHealth utilization among emerging adulthood including describing health self-management applications of mHealth and evaluating perceived effectiveness and ease of use across mHealth self-management applications. The findings will help inform the design and implementation of mHealth enabled prevention efforts in this critical population.

Materials and Methods

Study design and subjects

This data was collected from three different institutions of post-secondary education in a medium sized city located in the southeastern United States. The post-secondary educational institutions surveyed include a historically black college (HBCU), a R1 private university, and a technical community college. All three institutions collaborated on a community engaged and participatory research initiative to inform and improve population health management efforts. Each institution provided the research team with school e-mail addresses from either a random sample of their student body or the entire student body. Students received three e-mail prompts over two weeks from the research team to participate in the study and complete the survey through the Qualtrics web platform. After the survey was completed, students could enter a lottery to win a gift card to a popular online retailer. The study team had a response rate of 18%, which is typical of online surveys sent to students. Inclusion criteria required the respondents be currently enrolled at one of the participating institutions and have an active institutionally affiliated e-mail address. Respondents over the age of 30 or under the age of 18 were excluded from the analysis to focus on the emerging adulthood period of development. This study was approved by the Duke University Institutional Review Board.
Data collection

The survey captured a breadth of information from respondents including demographics, health behaviors, social determinants of health, and information on consumer/commercial oriented mHealth use. mHealth use was collected by asking respondents, “In the past year, how many health-related smartphone apps and/or wearable devices (example: dieting application, FitBit, Apple Watch, sleep tracking) have you used related to your health?”. Respondents provided information on age, gender, race, self-reported family income, and the degree they were pursuing. Data on daily fruit and vegetable in-take and moderate and vigorous exercise were collected using adapted items from the International Physical Activity Questionnaire. Health status was measured using a validated 1-item measure of self-reported health, Body Mass Index (BMI), a validated screening tool for clinical depression, and a self-reported item that asked respondents to indicate any current chronic condition diagnoses. Finally, primary care utilization was measured by asking respondents the frequency of getting routine outpatient care (“How long has it been since you last visited a healthcare provider for a routine checkup or yearly physical?”).

Respondents who indicated they used one or more apps/devices were asked to report what mHealth was used for by selecting a self-management category. Response options for self-management activities included: (1) to track how much activity/exercise I get (e.g. walking, running, cycling); (2) to help me watch what I eat (e.g. food intake, nutritional values); (3) to show or teach me exercises or workouts; (4) to track a health measure (e.g. heart rate, calories burned, blood pressure); (5) to help me manage a specific condition or behavior (e.g. diabetes, smoking cessation); (6) sleep tracking; (7) reproductive health (e.g. ovulation tracking, family planning, pregnancy progression); (8) stress/anxiety management, mindfulness, guided meditation. Respondents were asked additional questions about their experience using mHealth for each selected self-management activity. Of specific interest was self-reported ratings of ease of use and usefulness/effectiveness which have been shown to be predictive of intention to use, self-reported usage, and attitude toward technology use across a broad array of technologies.

Statistical analysis

A logit model was used to estimate the effect of included variables on the probability of using mHealth. The results are estimates of the average marginal effect of each variable on the likelihood of self-reported mHealth adoption. This approach was chosen given that the outcome (mHealth use) is binary and the strength of this dataset is the richness of covariates and diversity of the respondents. The primary hypothesis was higher BMIs would be associated with an increased likelihood of using mHealth as existing literature suggests that this association is present in the adult population.

The analysis also explored what self-management applications were supported by the use of mHealth. The descriptive analysis of the subgroup of emerging adults that reported using mHealth technology illustrates how these technologies were being used to manage health and promote specific health behaviors. Given existing literature examining how mHealth is used, the secondary hypothesis is that the two most common uses for mHealth technologies would be to improve nutrition and physical fitness.

Finally, descriptive statistics were calculated of perceived effectiveness and ease of use of mHealth technologies for each self-management activity to predict intention to use and sustained use. Ease of use and effectiveness ratings were analyzed to provide insights into the categories of mHealth enabled self-management activities that were the most likely to have an impact on long-term lifestyle modification.

Statistical analyses were performed by using Stata statistical software (version 15.1, StataCorp LLC, College Station, TX, USA).

Results

1,329 students were included in the analysis. A strength of this dataset is the diversity of respondents. 21.2% were African American, 40.5% White, 16.9% Asian, and 10.9% Hispanic (Table 1). Across the three post-secondary education institutions, 70.1% of the sample was female. The proportion of students that were overweight or obese was highest at the historically black college. The private university had the highest mean age and self-reported household income. mHealth use was prevalent and the difference in mHealth use was not statistically significant between the three institutions: private college (60.6%), technical community college (55.6%), and historically black college (56.4%).
Table 1. Description of population/analytic sample

<table>
<thead>
<tr>
<th></th>
<th>Private University</th>
<th>Technical Community College</th>
<th>Historically Black College</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>782</td>
<td>248</td>
<td>299</td>
<td>1,329</td>
</tr>
</tbody>
</table>

Demographics

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>505 (64.7%)</td>
<td>180 (73.2%)</td>
<td>246 (82.6%)</td>
<td>931 (70.1%)</td>
</tr>
<tr>
<td>≥ 24 years</td>
<td>373 (47.7%)</td>
<td>96 (38.7%)</td>
<td>88 (29.4%)</td>
<td>557 (41.9%)</td>
</tr>
<tr>
<td>Married or in a relationship</td>
<td>352 (45.5%)</td>
<td>101 (40.7%)</td>
<td>115 (38.9%)</td>
<td>568 (42.7%)</td>
</tr>
</tbody>
</table>

Race

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>395 (51.4%)</td>
<td>99 (39.9%)</td>
<td>44 (15.0%)</td>
<td>538 (40.5%)</td>
</tr>
<tr>
<td>African American</td>
<td>34 (4.4%)</td>
<td>54 (21.8%)</td>
<td>194 (66.0%)</td>
<td>282 (21.2%)</td>
</tr>
<tr>
<td>Asian</td>
<td>200 (26.0%)</td>
<td>17 (6.9%)</td>
<td>7 (2.4%)</td>
<td>224 (16.9%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>73 (9.5%)</td>
<td>55 (22.2%)</td>
<td>17 (5.8%)</td>
<td>145 (10.9%)</td>
</tr>
<tr>
<td>Other</td>
<td>66 (8.6%)</td>
<td>23 (9.3%)</td>
<td>32 (10.9%)</td>
<td>121 (9.1%)</td>
</tr>
</tbody>
</table>

Social Determinants

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocational or Associate’s degree</td>
<td>6 (0.8%)</td>
<td>206 (83.1%)</td>
<td>1 (0.3%)</td>
<td>213 (16.0%)</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>242 (30.9%)</td>
<td>18 (7.3%)</td>
<td>225 (75.3%)</td>
<td>485 (36.5%)</td>
</tr>
<tr>
<td>Graduate degree</td>
<td>520 (66.5%)</td>
<td>0 (0.0%)</td>
<td>66 (22.1%)</td>
<td>586 (44.1%)</td>
</tr>
</tbody>
</table>

Annual Income

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; $20,000</td>
<td>58 (7.8%)</td>
<td>42 (17.9%)</td>
<td>73 (25.6%)</td>
<td>173 (13.0%)</td>
</tr>
<tr>
<td>$20,000 – $49,999</td>
<td>106 (14.2%)</td>
<td>90 (38.3%)</td>
<td>98 (34.4%)</td>
<td>294 (22.1%)</td>
</tr>
<tr>
<td>$50,000 – $99,999</td>
<td>186 (24.9%)</td>
<td>64 (27.2%)</td>
<td>80 (28.1%)</td>
<td>330 (24.8%)</td>
</tr>
<tr>
<td>≥ $100,000</td>
<td>397 (53.1%)</td>
<td>39 (16.6%)</td>
<td>34 (11.9%)</td>
<td>470 (35.4%)</td>
</tr>
</tbody>
</table>

Health Status

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Overweight or obese</td>
<td>196 (25.1%)</td>
<td>110 (44.4%)</td>
<td>158 (52.8%)</td>
<td>464 (34.9%)</td>
</tr>
<tr>
<td>Any chronic illness</td>
<td>330 (42.2%)</td>
<td>139 (56.0%)</td>
<td>139 (46.5%)</td>
<td>608 (45.7%)</td>
</tr>
<tr>
<td>Depression</td>
<td>95 (12.1%)</td>
<td>58 (23.4%)</td>
<td>67 (22.4%)</td>
<td>220 (16.6%)</td>
</tr>
<tr>
<td>Excellent/Very Good Self-Reported Health</td>
<td>507 (64.8%)</td>
<td>124 (50.2%)</td>
<td>150 (50.3%)</td>
<td>781 (58.8%)</td>
</tr>
</tbody>
</table>

Health Behaviors

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mHealth use</td>
<td>472 (60.6%)</td>
<td>138 (55.6%)</td>
<td>168 (56.4%)</td>
<td>778 (58.5%)</td>
</tr>
<tr>
<td>PCP Visit within past year</td>
<td>572 (73.9%)</td>
<td>177 (71.7%)</td>
<td>233 (78.5%)</td>
<td>982 (73.9%)</td>
</tr>
<tr>
<td>Vegetable, Recommended daily intake</td>
<td>276 (35.5%)</td>
<td>78 (31.5%)</td>
<td>58 (19.4%)</td>
<td>412 (31.0%)</td>
</tr>
<tr>
<td>Fruit, Recommended daily intake</td>
<td>422 (54.0%)</td>
<td>126 (50.8%)</td>
<td>156 (52.2%)</td>
<td>704 (53.0%)</td>
</tr>
<tr>
<td>Moderate Exercise</td>
<td>202 (25.8%)</td>
<td>43 (17.3%)</td>
<td>56 (18.7%)</td>
<td>301 (22.6%)</td>
</tr>
<tr>
<td>Vigorous Exercise</td>
<td>95 (12.1%)</td>
<td>11 (4.4%)</td>
<td>21 (7.0%)</td>
<td>127 (9.6%)</td>
</tr>
</tbody>
</table>

A logit model was used to predict the likelihood of mHealth adoption based on the variables included in the model (Table 2). These results indicated that being overweight or obese is associated with a 10.5 percentage point increase in the predicted probability of using mHealth technology controlling for the other covariates included in the model. This finding offered support for the hypothesis that being overweight would be associated with a higher likelihood of using mHealth technologies.

Being female was associated with an over 14 percentage point increase in the predicted probability of using mHealth controlling for other variables in the model. Respondents from households under $20,000 were less likely to use mHealth when compared to respondents from households with incomes between $50,000-$99,999. Other income levels were not associated with a higher or lower probability of using mHealth. The endorsement of healthy behaviors also predicted mHealth use. Respondents that reported engaging in the recommended amount of moderate physical exercise according to the American Heart Association were more likely to use mHealth. In addition, eating the
The recommended daily intake of fruits is associated with a 6.2 percentage point increase in the likelihood of using mHealth.

Finally, there were associations between using mHealth across multiple measures of health status. Having a chronic illness was associated with using mHealth as was self-reported health of ‘Very Good’, or ‘Excellent’ which was associated with an 11 percentage point increase in the predicted probability of using mHealth. There was some preliminary evidence that suggests having screened positive for clinical depression is associated with a 6.6 percentage point decrease in the likelihood that a respondent used mHealth; however, this result was significant at a 90% confidence level. The other variables included in the model (vegetable consumption, relationship status, age, race, recommended vigorous exercise, annual primary care visit, or degree type) did not affect the likelihood of mHealth technology adoption.

Table 2. Logit model of marginal effects on mHealth adoption

<table>
<thead>
<tr>
<th>Variables</th>
<th>Logit Coefficients (SE)</th>
<th>Logit Marginal Effects (SE)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.655*** (0.138)</td>
<td>0.148*** (0.031)</td>
<td>[0.087, 0.208]</td>
</tr>
<tr>
<td>≥ 24 years</td>
<td>0.044 (0.148)</td>
<td>0.010 (0.033)</td>
<td>[-0.054, 0.074]</td>
</tr>
<tr>
<td>Married or in a relationship</td>
<td>0.151 (0.126)</td>
<td>0.033 (0.028)</td>
<td>[-0.021, 0.088]</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>-0.163 (0.197)</td>
<td>-0.036 (0.044)</td>
<td>[-0.123, 0.050]</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.041 (0.223)</td>
<td>0.009 (0.049)</td>
<td>[-0.087, 0.105]</td>
</tr>
<tr>
<td>Asian</td>
<td>-0.049 (0.186)</td>
<td>-0.011 (0.041)</td>
<td>[-0.092, 0.070]</td>
</tr>
<tr>
<td>Other</td>
<td>-0.074 (0.223)</td>
<td>-0.016 (0.050)</td>
<td>[-0.115, 0.082]</td>
</tr>
<tr>
<td><strong>Social Determinants</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocational or Associate’s degree</td>
<td>-0.054 (0.189)</td>
<td>-0.012 (0.042)</td>
<td>[-0.094, 0.070]</td>
</tr>
<tr>
<td>Graduate degree</td>
<td>0.118 (0.166)</td>
<td>0.026 (0.037)</td>
<td>[-0.046, 0.098]</td>
</tr>
<tr>
<td><strong>Annual Income</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; $20,000</td>
<td>-0.585** (0.205)</td>
<td>-0.133* (0.047)</td>
<td>[-0.225, -0.041]</td>
</tr>
<tr>
<td>$20,000 – $49,999</td>
<td>-0.031 (0.176)</td>
<td>-0.007 (0.039)</td>
<td>[-0.083, 0.069]</td>
</tr>
<tr>
<td>$50,000 – $99,999</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
</tr>
<tr>
<td>≥ $100,000</td>
<td>0.176 (0.164)</td>
<td>0.039 (0.036)</td>
<td>[-0.032, 0.110]</td>
</tr>
<tr>
<td><strong>Health Status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overweight or obese</td>
<td>0.484*** (0.138)</td>
<td>0.105*** (0.029)</td>
<td>[0.048, 0.163]</td>
</tr>
<tr>
<td>Any chronic illness</td>
<td>0.451** (0.131)</td>
<td>0.100*** (0.029)</td>
<td>[0.044, 0.156]</td>
</tr>
<tr>
<td>Depression</td>
<td>-0.295 (0.168)</td>
<td>-0.066 (0.038)</td>
<td>[-0.140, 0.008]</td>
</tr>
<tr>
<td>Excellent/Very Good Self-Reported Health</td>
<td>0.491*** (0.133)</td>
<td>0.110*** (0.030)</td>
<td>[0.052, 0.168]</td>
</tr>
<tr>
<td><strong>Health Behaviors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCP Visit within past year</td>
<td>-0.018 (0.142)</td>
<td>-0.004 (0.031)</td>
<td>[-0.065, 0.057]</td>
</tr>
<tr>
<td>Vegetable, Recommended daily intake</td>
<td>0.172 (0.144)</td>
<td>0.038 (0.031)</td>
<td>[-0.024, 0.100]</td>
</tr>
<tr>
<td>Fruit, Recommended daily intake</td>
<td>0.278* (0.128)</td>
<td>0.062* (0.028)</td>
<td>[0.006, 0.117]</td>
</tr>
<tr>
<td>Moderate Exercise</td>
<td>0.471** (0.163)</td>
<td>0.102** (0.034)</td>
<td>[0.035, 0.169]</td>
</tr>
<tr>
<td>Vigorous Exercise</td>
<td>-0.244 (0.234)</td>
<td>-0.054 (0.052)</td>
<td>[-0.157, 0.048]</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.006*** (.278)</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

Standard errors are presented in parentheses; Asterisks denote 5% (*) 1% (**) and .1% (***) significance; Pseudo $R^2 = 0.0683$; Log likelihood = -783.177

Respondents that reported using mHealth were asked follow up survey questions to capture how mHealth technologies were used and their perceived ease of use and effectiveness for each health self-management application (Tables 3 and 4).
Table 3. Effectiveness of mHealth technology by self-management activity

<table>
<thead>
<tr>
<th>mHealth Function</th>
<th>Prevalence</th>
<th>Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Not effective at all (5)</td>
</tr>
<tr>
<td>Activity and Exercise Tracking</td>
<td>645 (82.9%)</td>
<td>25 (3.9%)</td>
</tr>
<tr>
<td>Nutrition Tracking</td>
<td>311 (39.9%)</td>
<td>18 (5.8%)</td>
</tr>
<tr>
<td>Show Exercise and Workouts</td>
<td>276 (35.5%)</td>
<td>9 (3.3%)</td>
</tr>
<tr>
<td>Health Measure Tracking</td>
<td>273 (35.1%)</td>
<td>5 (1.8%)</td>
</tr>
<tr>
<td>Sleep Tracking</td>
<td>252 (32.4%)</td>
<td>10 (4.0%)</td>
</tr>
<tr>
<td>Manage Reproductive Health</td>
<td>151 (19.4%)</td>
<td>3 (2.0%)</td>
</tr>
<tr>
<td>Manage Stress or Anxiety</td>
<td>105 (13.5%)</td>
<td>11 (10.5%)</td>
</tr>
<tr>
<td>Manage Condition or Behavior</td>
<td>24 (3.1%)</td>
<td>2 (8.3%)</td>
</tr>
</tbody>
</table>

Table 4. Ease of use of mHealth technology by self-management activity

<table>
<thead>
<tr>
<th>mHealth Function</th>
<th>Prevalence</th>
<th>Ease of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Extremely difficult (5)</td>
</tr>
<tr>
<td>Activity and Exercise Tracking</td>
<td>646 (83.0%)</td>
<td>3 (0.5%)</td>
</tr>
<tr>
<td>Nutrition Tracking</td>
<td>313 (40.2%)</td>
<td>6 (1.9%)</td>
</tr>
<tr>
<td>Show Exercise and Workouts</td>
<td>275 (35.3%)</td>
<td>1 (0.4%)</td>
</tr>
<tr>
<td>Health Measure Tracking</td>
<td>274 (35.2%)</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td>Sleep Tracking</td>
<td>252 (32.4%)</td>
<td>1 (0.4%)</td>
</tr>
<tr>
<td>Manage Reproductive Health</td>
<td>152 (19.5%)</td>
<td>2 (1.3%)</td>
</tr>
<tr>
<td>Manage Stress or Anxiety</td>
<td>106 (13.6%)</td>
<td>1 (0.9%)</td>
</tr>
<tr>
<td>Manage Condition or Behavior</td>
<td>24 (3.1%)</td>
<td>0 (0.0%)</td>
</tr>
</tbody>
</table>

Physical activity and nutrition tracking were common self-management applications, lending support for the hypothesis that exercise and nutrition would be common uses of mHealth. However, despite the prevalence of mHealth technologies to assist with nutrition tracking, respondents indicated that the effectiveness for this function was lower than other self-management functions with 40.2% indicating nutrition tracking apps were extremely or very effective and only 28.1% reporting that they were extremely easy to use. Respondents, however, found mHealth was especially
useful for managing reproductive health with over 74.8% of respondents reporting that it was extremely or very effective and over 66.5% reporting that it was extremely easy to use. Respondents were less likely to find mHealth useful for managing stress or anxiety with only 35.2% reporting that mHealth was extremely or very effective. However, mHealth for managing stress or anxiety had relatively high ratings for ease of use with 83% of respondents indicating moderately or extremely easy to use.

The overall relationship of the ease of use and effectiveness ratings based on self-management is shown in Figure 1. mHealth self-management applications with higher effectiveness ratings also tended to have higher ease of use ratings. Ease of use and effectiveness ratings for each self-management category had a statistically significant positive correlation ranging from 0.29 to 0.47 with the exception of ‘Manage a Condition or Behavior’, likely due to the small sample size. For example, applications like reproductive health and health measure tracking had high levels of effectiveness and also had high levels of ease of use when compared to other applications. Similarly, nutrition tracking had lower levels of both effectiveness and ease of use. However, effectiveness tended to be rated much lower than ease of use. This was especially true for managing stress and anxiety, which had low levels of effectiveness but high levels of ease of use.

Discussion

Facilitating effective mHealth utilization within emerging adulthood is increasingly being recognized as an opportunity to promote health self-management and improve health outcomes later in life. As health care delivery models assume more responsibility for managing the health of defined populations and communities, novel ways that leverage available mHealth technologies to promote health self-management is a key consideration. This study confirms previous literature that indicates that younger adults are more likely to use mHealth technologies and provides insights into the determinants that predict mHealth adoption and the health self-management activities that these technologies support.

mHealth use was much higher in this sample of emerging adults (58.5%) than among a nationally representative sample of respondents found in relevant recent studies with mHealth usage rates of 22.8% in 2017 and 11.7% in 2012. The dramatic difference could be attributed to the focus on a younger population and the continuation of a general trend towards greater mHealth use. While mHealth use was ubiquitous among respondents, these findings suggest that a key barrier to mHealth adoption in emerging adulthood is related to socio-economic status. Participants from low income families (<$20,000) had a 13.3 percentage point lower likelihood of mHealth adoption when

Figure 1. mHealth effectiveness and ease of use ratings of self-management activity ordered by effectiveness
compared to middle income families ($50,000-$99,999). Previous research suggests that the burden of cost associated with mHealth and lower technology literacy levels are potential explanations for this finding.\textsuperscript{32}

Being overweight or obese and having any chronic illness is associated with higher likelihood of mHealth adoption. This could be a result of these individuals seeking technologies to manage their condition. This is consistent with a previous finding that mHealth users were more likely to be obese.\textsuperscript{5} These results also reveal that being in excellent or very good self-reported health was associated with an increased likelihood of mHealth adoption. This is notable given that self-reported health is predictive of health care expenditures and mortality.\textsuperscript{24}

Female emerging adult respondents were more likely to adopt mHealth technologies, a finding that has been replicated in older populations.\textsuperscript{19,20} A novel finding of this analysis suggests that being at risk of depression was associated with a lower likelihood of using mHealth, however, this result was only significant at a 90% confidence level, thus further research is required to confirm this relationship. Finally, this research provides evidence that there is a relationship between mHealth adoption and the endorsement of recommended beneficial health behaviors. Engaging in the recommended amount of moderate exercise and fruit intake were both associated with a higher likelihood of using mHealth. This is an area of particular promise and further research is required to evaluate the causal direction of this association and to explore the potential implications for designing more effective mHealth interventions.

This research is the first to describe the health self-management activities that mHealth technologies are used for and provide user feedback on perceived effectiveness and ease of use based on the technology acceptance model. Perceived effectiveness and ease of use are predictive of intention to use health related technologies, a key consideration given that user acceptance is important to enable health informatics adoption in health service delivery.\textsuperscript{5,18} Effectiveness and ease of use have been shown to be influenced by individual and community contextual factors such as gender, culture, social determinants, and demographic characteristics.\textsuperscript{17} It will be important to also understand how these determinants are associated with continued use, and more importantly, positive health outcomes. Overall, effectiveness ratings tended to be strikingly lower than ease of use ratings across all self-management activities. This illustrates the challenge of designing mHealth to be effective for behavior change, it requires more than the technology being accessible and user-friendly. It is important to note that the most commonly used self-management activities were not the activities that respondents indicated to be the highest levels of effectiveness and ease of use. Nutrition tracking, for example, was the second most prevalent self-management activity but had lower effectiveness and ease of use ratings when compared to other self-management applications. Conversely, reproductive health self-management had lower rates of mHealth adoption but had the highest effectiveness and ease of use ratings. A possible explanation for this phenomenon is that respondents found that passive mHealth technologies that tracked activity or a health measure to be both effective and easy to use when compared to a self-management activity like stress management and nutrition tracking which requires greater user engagement.

Future directions for this work could create a more specific “user profile” and risk stratification to target mHealth interventions to users that are receptive to this medium of clinical or public health interventions. To do so, special consideration will be paid to the design of mHealth enabled self-management interventions to maximize effectiveness and the user experience. As opportunities to implement mHealth enabled prevention efforts that target emerging adults are explored, researchers and practitioners alike must continue to evaluate opportunities to personalize mHealth interventions and delivery models to realize the full potential health benefits.

Limitations:

This work has several notable limitations. First, given the relatively modest, but not unexpected response rate, there should be caution when generalizing these results to all emerging adults in post-secondary institutions. As with many observational cross-sectional survey study designs, a self-selection bias could make the study sample systematically different from the population of interest. The respondents were all located in the same southeastern city which limits external validity to other geographic regions. The lack of longitudinal data limits the ability to determine the extent of sustained use of mHealth technologies and whether they fulfilled their intended outcome – a key consideration for mHealth to modify health behaviors.

Conclusion

This study provides important insights into mHealth use and their perceived ease of use and effectiveness by self-management application in emerging adulthood in diverse post-secondary institutions of learning. mHealth use is high among respondents when compared to the general population.\textsuperscript{5,20} Higher likelihood of mHealth adoption was
associated with being overweight or obese, female, engaging in moderate exercise, eating the recommended amount of daily fruit, having a chronic condition, and higher levels of self-reported health. In contrast, low household income and screening positive for depression risk (90% confidence level) was associated with a decreased likelihood of mHealth adoption. Tracking exercise and nutrition were the most common applications of mHealth. Ease of use and effectiveness were the highest for mHealth self-management applications associated with reproductive health and health measure tracking and the lowest for managing stress and anxiety. A trend that emerged in the data is that passive self-management functions received higher ease of use and effectiveness ratings. Ease of use and effectiveness ratings were largely correlated with ease of use being consistently rated higher than effectiveness across self-management activities. This finding suggests that an mHealth technology that is easy to use is necessary but not necessarily sufficient to have an impact on health behaviors. Next steps for this work should focus on opportunities to design mHealth interventions that are tailored to the preferences and needs of emerging adults to support health self-management during this critical developmental period.

Funding

This research was supported by the Duke Center for Personalized Health Care and by the Duke Clinical and Translational Science Award, NIH Award number UL1TR001117.

Acknowledgements

We gratefully acknowledge Cindy Mitchell for assistance preparing and editing this manuscript.

References

29. StataCorp L. Stata statistical software. College Station TX. 2009.
Perceived Usefulness and Acceptance of Communication Support System in Laparoscopic Surgery

Yuanyuan Feng, MS, Jatin Chhikara, MS, Jordan Ramsey, BS, Helena M. Mentis, MS PHD
University of Maryland Baltimore County, Baltimore, MD, USA

ABSTRACT

In laparoscopic surgery, senior surgeons spend great efforts to convey knowledge to their trainees. Any misinterpretation of the instructions may lead to inevitable errors that have a great impact on patient safety. To support efficient knowledge sharing, we design a communication support system, Virtual Pointer, to convey trainers’ gestures directly onto the laparoscopic video for trainees to see. We implement the design in an international surgical conference and evaluate the perceived usefulness and acceptance of the system among senior surgeons and surgical trainees. The results showed that the system facilitated the trainees to complete the task, increased their knowledge, and reduced the trainers’ guiding efforts. The study indicates that conveying the gestures directly onto video has the potential to support the intraoperative communication. Barriers in the acceptance of the system highlight the design of novel interaction mechanisms to unobtrusively integrate this technique into the surgical workflow.

INTRODUCTION

The operating room (OR) is the key place for residents to acquire surgical expertise. However, over half of surgical adverse events are attributable to the OR-based surgical training[1, 2]. The main cause for these errors is the communication failure in the high-stakes environment with multiple demands between residents and senior surgeons[3]. Residents, on one hand, need to efficiently follow the instructions given by their trainers. On the other hand, they master surgical skills through in-situ learning, which requires them to make their own decisions and get corrected and confirmed from the trainers[3]. Fulfilling these demands rely on the ability of expert surgeons in conveying the knowledge, that is in response to their trainees’ knowledge background, experience levels, and comprehensive abilities.

Previous research has shown that actions, such as pointing or instrument movements, have essential communicative functions that facilitate the conveying and understanding of knowledge[4, 5]. Expert surgeons often use actions to clarify the referred anatomical structure[6], manage the trajectory of the operation[7, 8], and demonstrate the technical skills[9]. These actions are sometimes transformed into explicit instructions when necessary in open surgery. For instance, an expert surgeon may draw on the skin to indicate the path of the incision and to mark the major landmarks of the anatomy[7]. The drawing specifies where the residents should attend and ensures that they are on the same page throughout the operation.

However, in laparoscopic surgery, where the direct view of the operative field limits the ability to directly gesture on the anatomical structures, expert surgeons take great effort in appropriating the view to guide their trainees to see the target structure, make meaning of it, and using it for their actions[3, 9, 10]. Expert surgeons often guide trainees’ hands to maneuver the laparoscopic camera, accompanied by a series of verbal explanations and gestures over the monitor to reveal the subtle changes in the structure to translate what they are seeing into the abstract anatomy in the textbook[3, 9]. The surgical instruments are used to point at the target or draw imaginary lines to elucidate the structures embedded in the tissues. Trainees, on the other hand, need to parse, envision and make sense of their trainers’ actions to perceive the locations and directions. This often engenders a series of checking, clarifying, aligning of the information that is already presented[10]. In this process, any misinterpretations of these actions may lead to inevitable technical and judgment errors that have a great impact on patient safety[11].

To overcome these problems, we hypothesized that conveying the gestures directly onto the video is beneficial. Previous research has shown that communication becomes more effective when annotation and gesturing tools are provided for the shared view of the workspace[12, 13]. For instance, Fussell et al. found that the pointing and representational gestures added on a shared view of a robot assembly task effectively facilitated task communication and reduced performance time[12]. Tang et al. projected the hands over the shared view and found that the gestures allowed the collaborators to be more engaged in communication and the tasks[13]. These studies focused on the communication in completing a task. The communication in OR-based surgical training, yet, entails not only instructions necessary in operating, but also the in-situ knowledge for the trainees to become competent in surgery[3, 9]. It is unknown whether the gesturing tool supports conveying expertise in surgical training.
We designed a Virtual Pointer system, which enables a trainer to point or draw a freehand sketch, i.e., telestrating, over a laparoscopic video for a trainee to see during the operation\textsuperscript{14}. We implement the design in an international surgical conference and evaluate the perceived usefulness and acceptance of the system among expert surgeons and residents. The perceived usefulness is defined as “the degree to which a person believes that using a particular system would enhance his or her job performance.”\textsuperscript{15} In our evaluation, we mainly focus on knowledge sharing in surgical training. The job performance here refers to trainees’ learning performance, i.e. the knowledge they gained, the skills and abilities they developed, and the effort they expended in training, as well as the trainers’ ability in conveying the knowledge. User acceptance is the degree to which a person is willing to use the system in their work practice\textsuperscript{15}. In our study, the user acceptance refers to the trainers’ willingness to use the system in conveying knowledge, as well as the trainees’ willingness to gain knowledge through the annotation or reference on the video provided by the trainers using the Virtual Pointer. The evaluation presents the surgeons and surgical trainees’ perspectives on how they are going to use a Virtual Pointer-like communication support system, what benefits they will gain from using the system, as well as the concerns and barriers in implementing the system in the surgical environment. Based on the findings, we discuss the implications for designing an intraoperative communication support system.

**METHODS**

**Overview of the Virtual Pointer system**

The Virtual Pointer system was designed aiming to facilitate the conveyance of expert knowledge by enabling trainers to point or draw on the laparoscopic video for the trainees to see during an operation. The program works by using a combination of audio key words and hand movements to trigger different functionality. The Microsoft Kinect is used as the gesture and voice control sensor.

![Virtual Pointer Interface](image)

**Figure 1.** The interface of the Virtual Pointer program\textsuperscript{14}. (1: the selected mode; 2: the audio commands; 3: the user’s movement feedback; 4: the laparoscopic view.)

Figure 1 shows the interface of the system. The collection of the verbal commands is shown in the upper left corner. The current mode is presented in the center above the laparoscopic view. The lower left corner shows the user’s skeleton to provide timely feedback of the user’s movement. To awaken the Kinect, the first command is verbally saying “Kinect ready”. When this is said the Kinect starts detecting other verbal cues and gestures. There are two verbal cues the Kinect is looking for, either “Kinect draw” or “Kinect point”, to switch between the drawing mode and the pointing mode. In the pointing mode, the user moves the hand to control a small green circle, which acts as a pointer. In the drawing mode, the user closes their hand to draw over the video. The position of the pointer and the drawing responds to the position of the user’s hand. To clear the screen of all annotations, there is the verbal command “Kinect clear”. When the program is finished being used, the voice command “Kinect close” can be used at any time to set the program to sleep and stop the Kinect from detecting.

**Study Procedure**

Virtual Pointer is a novel instructional system for surgery training. To minimize the learning curve, we first demonstrated how the system works and gave a brief introduction of the study. The surgeons or surgical trainees who consented to participate in the study completed a preliminary questionnaire on demographics and surgical experience.
Based on their experience level (attending surgeon, surgical fellows, and residents) the participants were assigned into two groups – the trainer group (attending surgeons) and the trainee group (surgical fellows and residents).

The trainers were asked to use the Virtual Pointer to guide the researcher to perform a simulated laparoscopic task. The researcher gained the fundamental knowledge of the task from the general surgery textbook and case videos. The trainees were asked to perform the task with the guidance from the researcher using the Virtual Pointer. A general guidance transcript was developed to include the features of pointing and drawing – the pointing was for the anatomical identification, and the drawing was for the indication of where to cut. The in-situ guidance was based on the general guidance transcript and adapted based on the needs of the trainees.

Afterward, the participants completed an evaluation questionnaire on their experience, outcomes expectations, attitudes and intentions of using the system. Due to the different demands in using the technology between the trainers and trainees, we administered two sets of questionnaires – one for the trainers on conveying the knowledge, and one for the trainees on acquiring the knowledge. The questionnaire was adapted from previous research on perceived usefulness and acceptance of educational technologies16,17,18, including 21-item 7-point Likert scale and open-ended questions. The study was approved by the University of Maryland, Baltimore County institutional review board. Verbal consent was obtained from all participants before their participation.

The Simulated Task

The simulated task in the experiment was to separate the cystic duct and cystic artery on a paper model. It is an important step in achieving the “critical view of safety” in the laparoscopic cholecystectomy procedure. This task requires the trainees not only to correctly identify the cystic duct and cystic artery that are embedded in the peritoneum, but also envision the imaginary line of safety over the two structures to avoid dissecting into the common bile duct. Thus, this task involves the effort of both referring to task objects or locations and managing the process of dissection.

Study Settings and Participants

The Virtual Pointer system was connected to a Fundamentals of Laparoscopic Surgery (FLS) Box Trainer System (Limbs & Things, Bristol, UK). The Microsoft Kinect sensor (Microsoft, Washington, USA) was oriented towards the trainer. The trainee stood in front of the FLS Box with the dominant hand controlling the scissors and the other hand manipulating the Maryland dissector. The Virtual Pointer provided an overlay of the reference and annotation from the trainer to the laparoscopic video for the trainee to see. As the transfer of the video created a 2-second lag, we presented the telestration on a secondary display (Figure 2).

Figure 2. Study settings in 2017 Society of American Gastrointestinal Endoscopic Surgeons (SAGES) Annual Meeting. (1: FLS Box Trainer System; 2: Laptop running the Virtual Pointer; 3: Microsoft Kinect sensor; 4: Primary display for the laparoscopic view; 5: Secondary display for the telestration; 6: Virtual Pointer Introduction.)

Recruitment

Attendees of the Telestration-Guided Laparoscopic Visual Skills Practice Station at the Learning Center during the 2017 Society of American Gastrointestinal Endoscopic Surgeons (SAGES) Annual Meeting in Houston, Texas, were
invited for participation. A total of 18 trainers (10) and trainees (8) participated in the study, including four female surgeons and 14 males. Experience levels ranged from novice medical students to senior surgeons with more than 20 years in practice. Among the 18 post-task questionnaires collected, 3 (1 trainer and 2 trainees) were identified to contain inconsistent answers for the same assertion asking in different directions and were removed from the analysis.

Data Analysis

We conducted non-parametric tests for the analysis of the questionnaire responses. For each question, we conducted the one-sample Wilcoxon signed-rank test comparing the sample median with the neutral rate. We used the Mann Whitney U test to compare the responses of the shared questions between the trainers and trainees. Spearmen’s correlations were calculated between questions. The median response of each question was reported, along with the p-values. The analysis was conducted in R version 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Perceived Support in Conveying Knowledge

Overall, 75% of the trainers were satisfied with the support of Virtual Pointer in conveying the knowledge during the task, but their ratings were insignificant (M = 5.5, p = 0.137) (Figure 3). 87.5% trainers agreed that the Virtual Pointer saved their efforts spent on teaching (M = 5.0, p = 0.037). However, only around a half of the trainers were satisfied with the amount of time in conveying knowledge via Virtual Pointer (M = 6.0, p = 0.101). And, 55.6% trainers thought the technology may increase their desire to teach (M = 5.0, p = 0.354). In the open-ended questions, the trainers questioned about the efficiency and reliability of the system. Trainer #5, who was not satisfied with the support of the Virtual Pointer, commented, “it requires I do not use the instrument while telestrating”.

Perceived Support in Acquiring Knowledge

Overall, both the trainers and trainees perceived the Virtual Pointer’s benefit in increasing trainees’ knowledge (Trainer: M = 5.0, p = 0.058; Trainee: M = 5.5, p = 0.041) (Figure 4). The most commonly perceived aspect of learning that the Virtual Pointer could improve was the procedure knowledge, with the ratings of both the trainers (M = 6.0, p = 0.031) and the trainees (M =5.5, p = 0.038) significantly greater than neutral. Although the trainers significantly agreed that the system could increase trainees’ knowledge in anatomy (M = 5.0, p = 0.026), the trainees found the limited support in their anatomical knowledge building (M = 5.0, p = 0.066). The system’s effect on laparoscopic skills was the least perceived by only 33.3% trainers and 66.7% trainees, with all the ratings prone to neutral. In addition to accumulating knowledge, the Virtual Pointer was perceived to have a positive effect on enabling trainees to apply the possessed knowledge into practices, by 77.8% trainers (M = 5.0, p = 0.046) and 83.3% trainees (M = 6.0, p = 0.039).

75.0% of the trainers found that the use of the Virtual Pointer may minimize the efforts that trainees spent on learning (M = 5.5, p = 0.036). 66.7% of the trainees agreed with the trainers that the technology reduced their learning efforts. Yet, their agreements were not significant (M = 5.0, p = 0.276). The reduced learning efforts may be due to the effect of the pointing or drawing on clarifying the target structure, as agreed by 83.3% trainees (M = 6.0, p = 0.041). Trainee #3 commented, “it was ease of use and clear”. Besides, 66.7% of the trainees indicated that the system encouraged them to learn more about the procedure (M = 6.0, p = 0.059).
Perceived Moderate Support in Task Performance

We asked the trainers and trainees to evaluate the effect of Virtual Pointer on their task performance, in terms of the easiness, time, effort and distraction (Figure 5). Overall, 66.7% of the trainers and trainees were satisfied with the support of Virtual Pointer in completing the task, however, their ratings were insignificant (Trainer: Median $= 5.0$, $p = 0.167$; Trainee: Median $= 5.5$, $p = 0.066$). 66.7% of the trainees thought that the Virtual Pointer made the task easier and reduced the amount of time to completion ($M = 5.5$, $p = 0.066$). However, the Virtual Pointer was perceived not effective in reducing the performance efforts ($M = 5.0$, $p = 0.144$). Trainee #6, who strongly agreed that it minimized the effort on performing the task, contributed this effect to the trainers, justifying that “it saves time for attendings to explain things to medical students and observers”. Interestingly, one-third of the trainees reported distractions and one-third of the trainees reported no distractions resulted from the Virtual Pointer ($M = 4.0$, $p = 0.705$). These distractions were caused by the trainees’ concern about the accuracy of the information provided in the pointing or drawing, as justified by Trainee #5 that “it is difficult to draw accurate lines and unreliable to follow voice commands and manual gestures”. Besides, as suggested by Trainee #3, the trainees had to switch their attention between the telestrated screen and the field of surgical view.

User Experience of the Virtual Pointer

Figure 4. Virtual Pointer in supporting acquiring knowledge. (AK: acquiring knowledge; The numbers in the circles show the medians of the ratings.)

Figure 5. Virtual Pointer in supporting task performance (TP: task performance; The numbers in the circle show the medians of the ratings.)
User experience is an important factor in user acceptance of technology and implementation of it in the workflow. Thus, we ask the trainers a couple of questions regarding the ease of use. As shown in Figure 6, more than half of the trainers had problems in using the system. Among them, 55.6% of the trainers thought that they would need the support of a technical person to be able to use the system. A main issue of the system, as identified by Trainer #1, was on the sensitivity in capturing the trainers’ hand movements, which interrupting their control of the system. Thus, the trainers relied on the immediate feedback from the system, as commented by Trainer #1, to verify the commands they gave to the system.

**Figure 6.** User experience of the virtual pointer in the surgical training. (UE: user experience; The numbers in the circle show the medians of the ratings.)

### Positive Attitude on the Virtual Pointer for Training

All trainers (M = 6.0, p = 0.007) and 66.7% trainees (M = 6.0, p = 0.063) agreed that using the Virtual Pointer in training was a good idea (Figure 7). 66.7% trainers (M = 6.0, p = 0.026) and trainees (M = 5.5, p = 0.059) agreed that Virtual Pointer was beneficial to trainees’ learning outcomes. Of all participants, only two trainers felt that using the system in training was unpleasant.

**Figure 7.** Attitude towards using the Virtual Pointer in surgical training. (A: attitude; The number in the circles show the medians of the ratings.)

### Intention to Use the Virtual Pointer

Although the trainers held a significantly positive attitude on the Virtual Pointer, they were conservative on using the Virtual Pointer in the lab (M = 5.0, p = 0.107), or the operating rooms (M = 5.0, p = 0.230). On the contrary, the trainees, who were moderately positive on the Virtual Pointer, would like to use the system in the lab training (M = 5.5, p = 0.026), but not in the operating rooms (M = 4.5, p = 0.257). Moreover, for using the system in varying difficulty levels, the approximate balance between ‘agree’ and ‘disagree’ indicates that case difficulty is not a decisive point for trainers and trainees to use the Virtual Pointer.

### Factors Associated with the Acceptance of Virtual Pointer

The Spearman’s correlation between the questions allows us to investigate the factors associated with the acceptance of the Virtual Pointer in surgical training. Mann Whitney U tests were conducted to assess any significant difference for any common questions shared between the trainers and trainees. No significant difference was found. Thus, we combined these two groups for the responses of the common questions in the correlation analysis. Besides, due to the
observed irrelevance of case difficulty for the use of Virtual Pointer, we excluded questions on the case difficulty (IE4, 5) and focused on the factors associated with the use in the lab and operating rooms (IE1, 2).

As shown in Table 1, the use of Virtual Pointer in the simulation lab (IE1) was significantly correlated with the support for task completion (TP1), acquired knowledge (AK1 – 5), and positive attitude on learning outcomes (A3). The use of the Virtual Pointer in the operating rooms was significantly correlated with the support for task completion (TP1), learning efforts (AK6) and acquired knowledge (AK1 – 3, 5). It is noteworthy that learning effort is significantly correlated with the use of the system in the operating rooms, while it is insignificantly correlated with the simulation lab. Task completion is more closely associated with the use of the system in the operating rooms, compared to the lab. On the contrary, acquired knowledge has a smaller effect on the use of the system in the operating rooms, compared to its high correlation with the use in the lab. Moreover, task completion (TP1) and learning efforts (AK6) were significantly correlated with the acquired knowledge, indicating that they may be the primary determinants of the usefulness of the system, as well as the acceptance of the system.

Table 1. Correlation between questions.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>TP1</th>
<th>AK1</th>
<th>AK2</th>
<th>AK3</th>
<th>AK4</th>
<th>AK5</th>
<th>AK6</th>
<th>A1</th>
<th>A2 #</th>
<th>A3</th>
<th>IE1</th>
<th>IE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP1</td>
<td>15</td>
<td>5.11</td>
<td>0.408</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK1</td>
<td>15</td>
<td>5.33</td>
<td>0.330</td>
<td>.744**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK2</td>
<td>15</td>
<td>5.25</td>
<td>0.300</td>
<td>.338**</td>
<td>.801**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK3</td>
<td>15</td>
<td>5.28</td>
<td>0.284</td>
<td>.418**</td>
<td>.739**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK4</td>
<td>15</td>
<td>5.61</td>
<td>0.307</td>
<td>.286</td>
<td>.465</td>
<td>.511</td>
<td>.1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK5</td>
<td>15</td>
<td>5.39</td>
<td>0.319</td>
<td>.598**</td>
<td>.760**</td>
<td>.765**</td>
<td>.715</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK6</td>
<td>14</td>
<td>5.04</td>
<td>0.406</td>
<td>.458</td>
<td>.813**</td>
<td>.753**</td>
<td>.313</td>
<td>.806**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>15</td>
<td>5.99</td>
<td>0.284</td>
<td>.483</td>
<td>.477</td>
<td>.513</td>
<td>.457</td>
<td>.052</td>
<td>.529**</td>
<td>.387</td>
<td>.1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>15</td>
<td>4.97</td>
<td>0.365</td>
<td>.015</td>
<td>.162</td>
<td>.140</td>
<td>.008</td>
<td>.012</td>
<td>.109</td>
<td>.041</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>15</td>
<td>5.31</td>
<td>0.287</td>
<td>.125</td>
<td>.423</td>
<td>.428</td>
<td>.384</td>
<td>.292</td>
<td>.604**</td>
<td>.530</td>
<td>.367</td>
<td>.160</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE1</td>
<td>14</td>
<td>5.35</td>
<td>0.339</td>
<td>.570**</td>
<td>.709**</td>
<td>.714**</td>
<td>.564**</td>
<td>.719**</td>
<td>.490</td>
<td>.255</td>
<td>.041</td>
<td>.612**</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>14</td>
<td>4.63</td>
<td>0.372</td>
<td>.808**</td>
<td>.737**</td>
<td>.641**</td>
<td>.788**</td>
<td>.378</td>
<td>.640**</td>
<td>.638**</td>
<td>.285</td>
<td>.005</td>
<td>.502</td>
<td>.696**</td>
<td>1.000</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).
# The scale is reversed.

The Spearman’s correlations between unique questions for either the trainers or the trainees were further analyzed. Table 2 shows the correlations between trainer questions. Easy to use was significantly correlated with the use of Virtual Pointer in the lab and in the operating rooms. The satisfactory on the support of the system for conveying knowledge is significantly correlated with the use of Virtual Pointer in the operating rooms, while insignificantly
correlated with the use in the lab. The increase of desire to teach is significantly correlated with the use in the lab, but insignificantly correlated with the use in the operating rooms.

Table 2. Correlation between unique questions for the trainers.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>CK1</th>
<th>CK2</th>
<th>CK3</th>
<th>CK4</th>
<th>UE1</th>
<th>UE2</th>
<th>IE1</th>
<th>IE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CK1</td>
<td>8</td>
<td>5.13</td>
<td>0.639</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CK2</td>
<td>9</td>
<td>4.89</td>
<td>0.539</td>
<td>.884*</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CK3</td>
<td>8</td>
<td>5.25</td>
<td>0.412</td>
<td>0.455</td>
<td>0.199</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CK4</td>
<td>9</td>
<td>4.56</td>
<td>0.556</td>
<td>0.148</td>
<td>0.199</td>
<td>0.423</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UE1</td>
<td>9</td>
<td>5.11</td>
<td>0.484</td>
<td>.790*</td>
<td>0.65</td>
<td>0.501</td>
<td>-0.544</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UE2</td>
<td>9</td>
<td>4.44</td>
<td>0.648</td>
<td>0.057</td>
<td>0.357</td>
<td>0.511</td>
<td>-0.415</td>
<td>0.115</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE1</td>
<td>8</td>
<td>4.88</td>
<td>0.479</td>
<td>0.689</td>
<td>0.199</td>
<td>0.637</td>
<td>.761*</td>
<td>.897*</td>
<td>-0.423</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>8</td>
<td>4.75</td>
<td>0.590</td>
<td>.847*</td>
<td>0.199</td>
<td>0.672</td>
<td>0.448</td>
<td>.951*</td>
<td>-0.228</td>
<td>.917*</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Among trainees’ unique questions, easier identification of target was significantly correlated with perceived task performance (Table 3). However, we have limited power to associate any factors to the acceptance of the system.

Table 3. Correlations between unique questions for the trainees.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>AK7</th>
<th>AK8</th>
<th>TP2</th>
<th>TP3</th>
<th>TP4</th>
<th>TP5</th>
<th>IE1</th>
<th>IE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK7</td>
<td>6</td>
<td>5.830</td>
<td>0.563</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK8</td>
<td>6</td>
<td>5.500</td>
<td>0.632</td>
<td>-0.079</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP2</td>
<td>6</td>
<td>5.500</td>
<td>0.477</td>
<td>.833*</td>
<td>0.143</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP3</td>
<td>6</td>
<td>5.500</td>
<td>0.500</td>
<td>.833*</td>
<td>0.143</td>
<td>1.000*</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP4</td>
<td>6</td>
<td>5.000</td>
<td>0.401</td>
<td>.894*</td>
<td>0</td>
<td>.909*</td>
<td>.909*</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP5</td>
<td>6</td>
<td>4.170</td>
<td>0.563</td>
<td>0.242</td>
<td>0.318</td>
<td>0.167</td>
<td>0.167</td>
<td>0.061</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE1</td>
<td>6</td>
<td>5.830</td>
<td>0.428</td>
<td>0.572</td>
<td>0.567</td>
<td>0.636</td>
<td>0.636</td>
<td>0.461</td>
<td>0.111</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>IE2</td>
<td>6</td>
<td>4.500</td>
<td>0.833</td>
<td>0.485</td>
<td>0.111</td>
<td>0.788</td>
<td>0.788</td>
<td>0.485</td>
<td>0.045</td>
<td>0.652</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

DISCUSSION

In this study, we evaluated the perceived usefulness and acceptance of the Virtual Pointer system, which was designed to support efficient knowledge sharing in surgical training, among senior surgeons and surgical trainees. The study suggests that conveying the gestures directly onto the laparoscopic video has the potential to support intraoperative communication and increase the efficiency of the performance. The participants liked the idea of directly pointing and drawing on the video. The Virtual Pointer was perceived to support the trainees efficiently identify the target referred by the trainers on the monitor, enhancing the trainees’ anatomical knowledge and their understanding of the procedure, minimizing the learning efforts, and facilitating the application of the knowledge into practice. The trainers found it easy to convey the knowledge with the system and could foresee the learning benefits their trainees would have in laparoscopic surgery with the improved mentoring strategy.

To date, extensive effort has been contributed in designing systems to support knowledge sharing in the healthcare environment, such as hand-off tools, information retrieving systems, and augmentative information display20-22. However, there have been limited studies examining the opinions and attitudes of users of such systems, or novel concepts in HIT23, even though user perceptions and preferences are essential in the success of system adoption21. Previous research tended to use the technology acceptance model (TAM) examining the users’ perceived usefulness16, 18. Our study suggests that although perceived usefulness from the users was an important factor, the root determinant is based on the perceptions of both the direct users and the stakeholders who would be affected by the system. For instance, in our study, the trainees are not the direct users of the system. Yet, their knowledge acquisition process is the main concern in the surgical training and in the trainers’ consideration of using the system in the lab or the OR. Besides, we found that the user experience of the system affects how the system is perceived to be used, when it would
be used, and the system efficacy. Responses in our open questions suggest that for a system to be implemented in the surgical environment, the main concern is on the accuracy of the information the system provides and the interruption it may cause into the surgeons’ workflow.

Different users even from similar professional backgrounds use the systems in different ways\textsuperscript{24}. In this study, we captured perceptual nuances of how the system would be used, what it would achieve, and where it could be used. This information could be used not only to validate technology objectives, but also to identify the target cases and environments for deployment, and to envision hindrances for implementation.

The different associated factors between the use in the lab and the use in the operating rooms reflect different system requirements in these two training environments. In the lab training, where trainees can gain proficiency in a safe environment, the system should support the initial learning, where a vast of factual knowledge is recalled and built up, motor skills are developed, and strategies are encouraged to be developed and tested out. In the operating rooms, where patient safety is emphasized, the system should target at improving the quality of performance, making the anatomical variants salient to minimize trainees’ effort in learning to see the field.

It is interesting to find that the trainees intended to use the system in the lab only, although the operating rooms are the key place for the trainees to gain expert vision on the anatomical structure and the decision making to proceed. Our correlation analysis between questions reveals several factors that hinder the acceptance of the system. The user experience is directly related to the success of the system adoption. However, the trainee indicated the concerns that the system may distract them from their task. The distraction may be due to that the illustration was shown on a secondary display, so that the trainees needed to frequently switch their attention between the laparoscopic video, which they were working on, and the Virtual Pointer screen, which they gained information from. This barrier highlights that the lag should be solved and the pointing and telestrating should be integrated into the trainees’ monitor. Besides, the trainers had problems in using the system. The trainers would need the support of a technical person to be able to use the system. They found the gestural and voice control was not very reliable. This barrier could be solved by providing a training session, where the system could be tested out by the users and further configured to recognize the gesture and speech of the individual surgeon. In addition, the system may interrupt the operation as the system required the trainer to pause the procedure and put down the instruments, waving in the air to draw any annotation. To address this issue, new interaction mechanisms that do not depend on arm movements can be designed to control the system. Furthermore, the system lacks precision. The trainees must spend extra effort to identify inaccurate annotations on the monitor, discuss them with the trainer, and make judgments on how to proceed. This suggests the algorithm to be improved to accurately relate the hand position to the pointer.

In this study, we evaluated the system in the Society of American Gastrointestinal Endoscopic Surgeons (SAGES) Annual Meeting, where most attendees are American surgeons. The strength of this setting is that it provided us a diverse pool of American surgeons from different hospitals. However, we acknowledge that surgical training varies among different countries. For example, residents may enter the OR at different levels or they may be trained using different tools. These factors may influence the efficacy of the system in surgical training. Besides, although, we have a rich collection of the trainers, who are the targeted user of the system, we have a limited sample size of the trainees. However, we identified the trend of perceived usefulness in the trainees’ responses and the trend was verified by the trainers’ perceived usefulness.

CONCLUSION

To facilitate the efficient conveyance of expert knowledge in laparoscopic training, we hypothesized to convey the gestures onto the video and designed the Virtual Pointer system that enables trainers to point or draw freehand sketches over a video for a trainee to see. In this study, we evaluated the perceived usefulness and acceptance of the system in surgical training. Our study indicates that conveying the gestures directly onto video is an effective and useful approach for the communication between surgeons. Barriers in the acceptance of the system highlight the design of novel interaction mechanisms to unobtrusively integrate this technique into the surgical workflow.

Acknowledgements

The authors gratefully acknowledge SAGES for equipment support. This work was partially supported by NSF Grants IIS #1422671 and #1552837. The authors thank Hannah McGowan for editing the manuscript, and the SAGES conference committee for the support of surgical instruments, training equipment, and the space.

References
15. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly. 1989 Sep 1;319:40.
Health Tracking and Information Sharing in the Patient-Centered Era: A Health Information National Trends Survey (HINTS) Study

Ashley C. Griffin, MSPH1,2, Arlene E. Chung, MD, MHA, MMCi1-3
1University of North Carolina (UNC) at Chapel Hill, Chapel Hill, NC, USA; 2Carolina Health Informatics Program, Chapel Hill, NC, USA; 3Department of Medicine & Department of Pediatrics & Program on Health and Clinical Informatics, UNC School of Medicine, Chapel Hill, NC, USA

Abstract

We examined the current state of digital health tracking and information sharing with health professionals among patients with chronic conditions using data from the National Cancer Institute’s 2018 Health Information National Trends Survey (HINTS). Descriptive statistics were used to examine the characteristics of health tracking and information sharing. Chi-squared tests were used to compare across groups, and multivariate logistic regression models were used to control for covariates. Between 17.4-37.6% of respondents reported sharing information with a health professional through either e-mail, monitoring device, text message, or online medical record message. There were sociodemographic differences across health tracking and information sharing modalities, and patients with chronic conditions disproportionately lacked Internet access, a basic cell phone, smartphone, or tablet compared to those without chronic conditions (p<0.05). This suggests there are sociodemographic and technology-based disparities for health tracking and information sharing for patients with chronic conditions.

Introduction

There is growing use of mobile health (mHealth) technologies, including smartphones, apps, wearable devices, and remote monitoring devices, to capture health-related data. Given the near ubiquity of mobile phones, with 95% of U.S. adults owning a mobile phone and 77% owning a smartphone in 20181, digital approaches are a promising way for patients to monitor their health due to their portability and convenience. The ability for patients to access, track, and share health information also provides opportunities for caregivers and care teams to longitudinally monitor behaviors and for interacting with the health system. Sharing health information with the care team facilitates patient-centered care through increased awareness, improved communication, and the ability to capture and transmit patient-generated data to incorporate into clinical care2. Further, the secondary use of patient-generated health data (PGHD) has been valuable in accelerating clinical, public health, and research insights3-5.

There are a number of national initiatives that have set a high priority for promoting data access and sharing for patients. The Centers for Medicare & Medicaid Services Meaningful Use Program (recently renamed the Promoting Interoperability Program) requires eligible professionals and health systems to provide increasing amounts of data access to patients, such as the ability to view, download, or transmit health information to a third party, and to use secure electronic messaging to communicate with health professionals (Stage 1 and 2)6. In 2019, patients must have access to their health information using a third-party app of choice through an Application Programming Interface (API)7. In February 2019, the Office of the National Coordinator for Health Information Technology also proposed a new rule to support individuals in the secure access, exchange, and use of their health information electronically8 to facilitate healthcare system adoption of APIs. This proposed rule discourages information blocking by healthcare systems, payers, and vendors as detailed in the 21st Century Cures Act8. Progress has also been accelerated by the release of an API built on the Fast Healthcare Interoperability Resources (FHIR) standard by the Health Level Seven (HL7) International Argonaut Project9, which allows interoperability between electronic health records and apps, such as Apple HealthKit10. Additionally, the Centers for Medicare & Medicaid Services proposed policy changes to the MyHealthEData11 and Blue Button 2.012 initiatives to improve access, data exchange, and care coordination for beneficiaries through APIs and trusted exchange networks13. Furthermore, research programs, such as the National Institutes of Health All of Us Research Program, which seeks to create a national research resource to study how individual differences in biology, environment, and lifestyle factors influence health, relies on individuals to share information about their health using APIs and other methods14.

Being able to leverage patient health data for clinical care relies on a patient’s access, ability, and willingness to collect and share digital data with health systems15. Several sociodemographic characteristics influence access and use of technologies, such as educational attainment, age, race/ethnicity, health status, and Internet access16. Understanding

1041
the characteristics that could influence digital health tracking and information sharing may allow health professionals to better understand who is likely to share PGHD and identify those who could benefit the most from sharing data. While not all patients will equally benefit from sharing PGHD (e.g., healthy individuals or those with a well-managed condition), data sharing may be particularly important for patients with chronic conditions as digital technologies can facilitate communication, real-time monitoring, or tailored coaching from the care team\(^2\). Patients with one or more chronic conditions may also have a greater need for health tracking and sharing health information with healthcare professionals\(^{17-20}\). While there is evidence that health tracking and data sharing may vary across sociodemographic factors, the latest HINTS survey data from 2018 allows the opportunity to extend prior research by expanding on how different communication modalities are associated with sociodemographic characteristics and chronic conditions\(^{20,21}\). Thus, the objectives of this study were to: 1) describe the current state of health tracking and health information sharing across sociodemographic variables and the number of chronic conditions, and 2) examine the relationship of various modalities through which one can share health information with health professionals among patients with chronic conditions in a nationally representative sample.

**Methods**

Data were from the National Cancer Institute’s 2018 Health Information National Trends Survey (HINTS) 5 Cycle 2, administered January to May 2018\(^{22}\). HINTS is a survey administered to a nationally representative sample of U.S. adults to measure how individuals access and utilize health information\(^{22}\). Survey questions related to health tracking were selected from the larger survey data set including: 1) “Has your tablet or smartphone helped you track progress on a health-related goal such as quitting smoking, losing weight, or increasing physical activity?”; 2) “Other than a tablet or smartphone, have you used an electronic device to monitor or track your health within the last 12 months? Examples include Fitbit, blood glucose meters, and blood pressure monitors”; 3) “In the past 12 months, have you used your online medical record to securely message health care professional within the last 12 months?”. Question 1-3, responses included “Yes” or “No.” For question 4, responses included “0”, “1 to 2 times”, “3 to 5 times”, “6 to 9 times”, or “10 or more times”\(^{22}\). Survey questions related to health information sharing included the following: 1) “In the past 12 months, have you used a computer, smartphone, or other electronic means to e-mail or [use] the Internet to communicate with a doctor or a doctor’s office?”; 2) “Have you shared health information from either an electronic monitoring device or smartphone with a health professional within the last 12 months?”; 3) “Have you sent a text message to or received a text message from a doctor or other health care professional within the last 12 months?”; 4) “In the past 12 months, have you used your online medical record to securely message health care provider and staff (for example, e-mail)”? Response options for these health information sharing questions were: “Yes”, “No”, “Don’t Know”, or “Not applicable.” “Don’t Know” and “Not applicable” responses were categorized as “No”\(^{22}\). For chronic conditions, respondents were asked whether a healthcare professional had ever diagnosed them with the following: diabetes or high blood sugar; high blood pressure or hypertension; a heart condition such as a heart attack, angina, or congestive heart failure; chronic lung condition, asthma, emphysema, or chronic bronchitis; arthritis or rheumatism; depression or anxiety disorder; or cancer.

All analyses were conducted using survey weighting and jackknife variance estimations provided by HINTS to be nationally representative of the U.S. adult population\(^{22}\). All analyses were conducted using SAS (version 9.4, Cary, NC, U.S.). A variable was created for the number of conditions (0 conditions, 1-2 conditions, or 3+ conditions). Descriptive statistics were used to show the characteristics of health tracking and health information sharing. Chi-squared tests were used to compare across groups of patients with 0, 1-2, and 3 or more chronic conditions. To examine the relationship among patients with chronic condition(s) and various information sharing modalities, four multivariable logistic regression models were used to control for relevant covariates with each model estimating the information sharing modality (e-mail, text message, secure online medical record message, and electronic monitoring device). The multivariate models to estimate information sharing via e-mail, text message, and online medical record message only included respondents who had Internet access, at least a basic cell phone, or had ever been offered access to the medical record by a health care provider or health insurer respectively. In the model to estimate information sharing through an electronic monitoring device or smartphone, all respondents were included because there was no survey question that specifically asked about ownership of an electronic monitoring device or smartphone. All models were adjusted for the following covariates: gender, age, race/ethnicity, region, education, household income, occupation, marital status, time since most recent check-up visit, caregiver status, health insurance status, and chronic conditions.
Results

The final analytic sample was comprised of 2,439 respondents after excluding 1,065 respondents who had incomplete sociodemographic data. To compare differences between included and excluded cases, we conducted a bias analysis which showed that excluded respondents may have been less diverse than the general population and were more likely to be older and female (p<0.05). The sample had 48.33% females and a mean age of 53.02 years with a SD of 19.01 years (see Table 1). Respondents were primarily Caucasian, less than 65 years old, resided in an urban location, completed some college or attained a college degree or more, had a household income over $75,000 per year, visited a doctor for a routine check-up within the past year, and had one or more chronic conditions. The majority of respondents (61.09%) accessed the Internet through multiple networks (e.g., Wi-Fi, cellular, etc.), but 12.25% reported that they did not have access to the Internet. Over half of participants (56.42%) had multiple electronic devices (e.g., smartphone, tablet, etc.). There were 3.60% of respondents that reported that they did not have any electronic devices and 7.36% had only a basic cell phone. Age, race/ethnicity, education, income, occupation, time since most recent check-up visit, being a caregiver, having health insurance, Internet access, and electronic device use were associated with the number of chronic conditions (p<0.05).

Table 1. Sample characteristics across chronic conditions

<table>
<thead>
<tr>
<th>Characteristics (%)</th>
<th>Total Overall n=2,439</th>
<th>0 Chronic Conditions n=702</th>
<th>1-2 Chronic Conditions n=1,186</th>
<th>3+ Chronic Conditions n=551</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>48.33</td>
<td>45.53</td>
<td>48.23</td>
<td>55.76</td>
</tr>
<tr>
<td>Male</td>
<td>51.67</td>
<td>54.47</td>
<td>51.77</td>
<td>44.24</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-34</td>
<td>26.32</td>
<td>40.35**</td>
<td>21.56**</td>
<td>5.41**</td>
</tr>
<tr>
<td>35-49</td>
<td>28.67</td>
<td>32.72**</td>
<td>27.76**</td>
<td>21.22**</td>
</tr>
<tr>
<td>50-64</td>
<td>29.40</td>
<td>22.19**</td>
<td>33.56**</td>
<td>34.75**</td>
</tr>
<tr>
<td>65+</td>
<td>15.61</td>
<td>4.74**</td>
<td>17.12**</td>
<td>38.62**</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>65.37</td>
<td>59.26*</td>
<td>70.00*</td>
<td>66.52*</td>
</tr>
<tr>
<td>African American</td>
<td>10.53</td>
<td>10.26*</td>
<td>9.81*</td>
<td>13.44*</td>
</tr>
<tr>
<td>Hispanic</td>
<td>15.61</td>
<td>19.66*</td>
<td>12.49*</td>
<td>15.02*</td>
</tr>
<tr>
<td>Other</td>
<td>8.49</td>
<td>10.82*</td>
<td>7.70*</td>
<td>5.02*</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>98.79</td>
<td>98.28</td>
<td>99.30</td>
<td>98.49</td>
</tr>
<tr>
<td>Rural</td>
<td>1.21</td>
<td>1.72</td>
<td>0.70</td>
<td>1.51</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School or Less</td>
<td>28.59</td>
<td>26.37**</td>
<td>27.28**</td>
<td>38.34**</td>
</tr>
<tr>
<td>Some College</td>
<td>40.43</td>
<td>40.06**</td>
<td>40.49**</td>
<td>41.17**</td>
</tr>
<tr>
<td>College Degree or More</td>
<td>30.98</td>
<td>33.57**</td>
<td>32.23**</td>
<td>20.49**</td>
</tr>
<tr>
<td>Household Income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; $20,000</td>
<td>16.60</td>
<td>14.90**</td>
<td>14.54**</td>
<td>27.36**</td>
</tr>
<tr>
<td>$20,000 - $34,999</td>
<td>11.27</td>
<td>10.65**</td>
<td>9.20**</td>
<td>19.24**</td>
</tr>
<tr>
<td>$35,000 - $49,999</td>
<td>12.35</td>
<td>8.51**</td>
<td>14.57**</td>
<td>15.22**</td>
</tr>
<tr>
<td>$50,000 - $74,999</td>
<td>18.22</td>
<td>16.99**</td>
<td>19.91**</td>
<td>16.11**</td>
</tr>
<tr>
<td>&gt; $75,000</td>
<td>41.56</td>
<td>48.95**</td>
<td>41.78**</td>
<td>22.07**</td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>61.29</td>
<td>70.66**</td>
<td>62.26**</td>
<td>34.34**</td>
</tr>
<tr>
<td>Unemployed</td>
<td>10.36</td>
<td>14.34**</td>
<td>6.92**</td>
<td>10.94**</td>
</tr>
<tr>
<td>Disabled</td>
<td>5.69</td>
<td>0.83**</td>
<td>5.40**</td>
<td>19.01**</td>
</tr>
<tr>
<td>Retired</td>
<td>16.12</td>
<td>5.40**</td>
<td>18.67**</td>
<td>35.48**</td>
</tr>
<tr>
<td>Other</td>
<td>6.54</td>
<td>8.77**</td>
<td>6.75**</td>
<td>0.23**</td>
</tr>
</tbody>
</table>
Table 1. Sample characteristics across chronic conditions, continued.

<table>
<thead>
<tr>
<th>Marital Status</th>
<th>Single</th>
<th>Married</th>
<th>Single</th>
<th>Married</th>
<th>Single</th>
<th>Married</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>46.00</td>
<td>54.00</td>
<td>47.93</td>
<td>52.07</td>
<td>42.90</td>
<td>57.10</td>
</tr>
</tbody>
</table>

| Most Recent Check-up    | 64.26  | 51.59** | 69.64**| 79.74** | 47.90  | 52.07** |
|                        | 18.22  | 23.20** | 15.37**| 14.38** | 23.20**| 25.20** |
|                        | 17.52  | 25.21** | 14.99**| 5.88**  | 14.38**| 14.99** |

| Caregiver (non-professional) | 15.35 | 13.16* | 15.47* | 20.56* | 84.65 | 86.84* |
|                             | 84.65 | 86.84* | 84.53* | 79.44* | 13.14*| 13.60* |

| Health Insurance          | 91.25 | 86.86**| 94.30**| 92.91**|
|                          | 8.75  | 13.14**| 5.70** | 7.09** |

| Internet Access           | 14.63 | 13.80* | 15.01* | 15.54* | 7.86  | 4.79* |
|                         | 3.98  | 4.43*  | 3.33*  | 4.79*  | 6.16  | 3.60* |
|                         | 7.68  | 5.93*  | 8.74*  | 8.91*  | 5.93  | 3.60* |
|                         | 0.37  | 0.44*  | 0.15*  | 0.85*  | 0.21  | 0.13* |
|                         | 61.09 | 65.85* | 60.54* | 50.70* | 37.56 | 38.02* |
| Multiple Internet networks | 12.25 | 9.55*  | 12.23* | 19.21* | 12.25 | 9.55* |

| Device Use               | 28.04 | 28.14**| 30.88**| 18.90**|
|                         | 4.58  | 3.76** | 4.46** | 7.04** |
|                         | 7.36  | 3.37** | 6.58** | 19.98**|
|                         | 56.42 | 61.12**| 55.08**| 48.64**|
| No devices              | 3.60  | 3.61** | 3.00** | 5.44** |

1Chi-squared tests were conducted across the chronic disease groupings.
*p < 0.05, **p < 0.01

Health Tracking and Health Information Sharing Characteristics. Health tracking and health information sharing proportions varied across respondents with 0, 1-2, and 3 or more chronic conditions (see Table 2). For health tracking, 43.16% of respondents used a smartphone or tablet to track progress on a health-related goal such as losing weight, quitting smoking, or increasing physical activity. Approximately 11% reported not having a smartphone or tablet. Other than a smartphone or tablet, 36.74% of respondents used an electronic device, such as a blood pressure monitor, blood glucose meter, or Fitbit, to track their health within the past 12 months. In terms of accessing online medical records (OMR) at least once in the past 12 months, 31.67% of respondents indicated that they had done so, however only 17.35% downloaded it to a computer or mobile device. Using an electronic device, smartphone, or tablet to track health and accessing or downloading the OMR onto a computer or mobile device were associated with the number of chronic conditions (p<0.05).

For health information sharing, the most common modality for sharing was e-mail or via the Internet with 37.61% using e-mail or the Internet to communicate with a health professional within the past 12 months. 30.73% of respondents sent/received a text message to/from a health professional, but 3.60% did not have a basic cell phone. There were 17.63% of respondents who shared health information from an electronic monitoring device or smartphone with a health professional. While 17.35% sent a secure message to a health professional through their OMR, nearly half of participants (46.62%) did not have access to their OMR. E-mailing, secure messaging through the OMR, and sharing information from an electronic monitoring device were all associated with the number of chronic conditions (p<0.05).
Table 2. Health tracking and information sharing characteristics across chronic conditions.

<table>
<thead>
<tr>
<th>Characteristics (%)</th>
<th>Overall n=2,439</th>
<th>0 Chronic Conditions n=702</th>
<th>1-2 Chronic Conditions n=1,186</th>
<th>3+ Chronic Conditions n=551</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Health Tracking</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Used a smartphone or tablet to track health-related goal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>43.16</td>
<td>47.67**</td>
<td>43.99**</td>
<td>29.09**</td>
</tr>
<tr>
<td>No</td>
<td>45.88</td>
<td>45.35**</td>
<td>46.43**</td>
<td>45.49**</td>
</tr>
<tr>
<td>Did not have smartphone/tablet</td>
<td>10.96</td>
<td>6.98**</td>
<td>9.58**</td>
<td>25.42**</td>
</tr>
<tr>
<td>Used an electronic device to track health</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>36.74</td>
<td>29.95*</td>
<td>39.92*</td>
<td>44.14*</td>
</tr>
<tr>
<td>No</td>
<td>63.26</td>
<td>70.05*</td>
<td>60.08*</td>
<td>55.86*</td>
</tr>
<tr>
<td>Accessed Online Medical Record (OMR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>21.71</td>
<td>20.25*</td>
<td>21.25*</td>
<td>26.88*</td>
</tr>
<tr>
<td>1-5</td>
<td>25.57</td>
<td>21.85*</td>
<td>29.20*</td>
<td>23.74*</td>
</tr>
<tr>
<td>6+</td>
<td>6.10</td>
<td>2.94*</td>
<td>8.05*</td>
<td>8.02*</td>
</tr>
<tr>
<td>Did not have access to OMR</td>
<td>46.62</td>
<td>54.96*</td>
<td>41.50*</td>
<td>41.36*</td>
</tr>
<tr>
<td>Downloaded OMR onto computer or mobile device</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>17.35</td>
<td>11.76*</td>
<td>21.93*</td>
<td>17.30*</td>
</tr>
<tr>
<td>No</td>
<td>36.03</td>
<td>33.28*</td>
<td>36.57*</td>
<td>41.34*</td>
</tr>
<tr>
<td>Did not have access to OMR</td>
<td>46.62</td>
<td>54.96*</td>
<td>41.50*</td>
<td>41.36*</td>
</tr>
<tr>
<td><strong>Health Information Sharing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Used e-mail or Internet to communicate with a health professional</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>37.61</td>
<td>31.19**</td>
<td>43.95**</td>
<td>34.19**</td>
</tr>
<tr>
<td>No</td>
<td>50.14</td>
<td>59.26**</td>
<td>43.82**</td>
<td>46.60**</td>
</tr>
<tr>
<td>Did not have Internet access</td>
<td>12.25</td>
<td>9.55**</td>
<td>12.23**</td>
<td>19.21**</td>
</tr>
<tr>
<td>Sent or received a text message to/from a health professional</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>30.73</td>
<td>26.15</td>
<td>33.91</td>
<td>33.22</td>
</tr>
<tr>
<td>No</td>
<td>65.67</td>
<td>70.24</td>
<td>63.14</td>
<td>62.03</td>
</tr>
<tr>
<td>Did not have a basic cell phone</td>
<td>3.60</td>
<td>3.61</td>
<td>2.95</td>
<td>4.75</td>
</tr>
<tr>
<td>Shared health information from an electronic monitoring device or smartphone with a health professional</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>17.63</td>
<td>11.15**</td>
<td>20.29**</td>
<td>25.88**</td>
</tr>
<tr>
<td>No</td>
<td>82.37</td>
<td>88.85**</td>
<td>79.71**</td>
<td>74.12**</td>
</tr>
<tr>
<td>Messaged health professional through OMR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>17.35</td>
<td>11.76*</td>
<td>21.93*</td>
<td>17.30*</td>
</tr>
<tr>
<td>No</td>
<td>36.03</td>
<td>33.28*</td>
<td>36.57*</td>
<td>41.34*</td>
</tr>
<tr>
<td>Did not have access to OMR</td>
<td>46.62</td>
<td>54.96*</td>
<td>41.50*</td>
<td>41.36*</td>
</tr>
</tbody>
</table>

1Chi-squared tests were conducted across the chronic disease groupings.
2Occurred within the past 12 months
*p < 0.05, **p < 0.01

Health Information Sharing Across Modalities. Sociodemographic characteristics of respondents varied across modalities of health information sharing through e-mail, text message, OMR message, or an electronic monitoring device/smartphone (see Table 3). Across all modalities, individuals with one or more chronic conditions had at least 1.50 or greater odds of sharing health information with a health professional compared to those without chronic conditions (p<0.05). Notably, respondents with three or more chronic conditions had 3.61 higher odds of sharing...
health information with an electronic device or smartphone with a health professional compared to those without chronic conditions (p<0.01).

Among respondents with Internet access, those who were male, 65 years and older, or had their most recent check-up three or more years ago had lower odds of e-mailing a health professional. Those who had at least a college degree, household income >$75,000, or had more than one chronic condition had higher odds of e-mailing a health professional. For text messaging with a health professional, those who were male, disabled, or without health insurance had lower odds of texting, while those who had more than one chronic condition had higher odds of texting. In terms of using an OMR for messaging, those who had one or more chronic condition had greater odds of doing so than those without chronic conditions. For sharing information using electronic devices or smartphones, those who were 50-64 years old or those who had their most recent check-up greater than a year ago had lower odds of sharing. Those who were African American, made $50,000 or more, or had more than one chronic condition also had higher odds of sharing information.

Table 3. Odds ratios (SE) of health information sharing across modalities during the past year.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>E-mail or Internet</th>
<th>Text Message</th>
<th>OMR Message</th>
<th>Electronic Device or Smartphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Referent: Female)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.71 (0.13)*</td>
<td>0.68 (0.18)*</td>
<td>0.97 (0.24)</td>
<td>0.81 (0.18)</td>
</tr>
<tr>
<td>Age (Referent: 18-34)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35-49</td>
<td>1.15 (0.24)</td>
<td>1.13 (0.23)</td>
<td>1.05 (0.34)</td>
<td>0.79 (0.29)</td>
</tr>
<tr>
<td>50-64</td>
<td>0.83 (0.21)</td>
<td>1.08 (0.22)</td>
<td>0.76 (0.29)</td>
<td>0.58 (0.27)*</td>
</tr>
<tr>
<td>65+</td>
<td>0.57 (0.28)*</td>
<td>0.82 (0.35)</td>
<td>0.75 (0.38)</td>
<td>0.80 (0.31)</td>
</tr>
<tr>
<td>Race/Ethnicity (Referent: Caucasian)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>1.02 (0.22)</td>
<td>0.90 (0.29)</td>
<td>1.12 (0.35)</td>
<td>1.83 (0.30)*</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.71 (0.22)</td>
<td>0.91 (0.20)</td>
<td>0.88 (0.28)</td>
<td>0.83 (0.26)</td>
</tr>
<tr>
<td>Other</td>
<td>1.65 (0.42)</td>
<td>1.33 (0.26)</td>
<td>1.09 (0.40)</td>
<td>1.17 (0.31)</td>
</tr>
<tr>
<td>Region (Referent: Urban)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>0.97 (0.94)</td>
<td>1.10 (0.74)</td>
<td>1.51 (0.89)</td>
<td>0.95 (0.79)</td>
</tr>
<tr>
<td>Education (Referent: Some College)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School or Less</td>
<td>0.73 (0.23)</td>
<td>0.68 (0.21)</td>
<td>0.93 (0.37)</td>
<td>0.72 (0.22)</td>
</tr>
<tr>
<td>College Degree or More</td>
<td>1.60 (0.17)**</td>
<td>1.26 (0.19)</td>
<td>1.25 (0.30)</td>
<td>0.91 (0.18)</td>
</tr>
<tr>
<td>Household Income (Referent: &lt; $20,000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$20,000 - $34,999</td>
<td>1.15 (0.33)</td>
<td>0.79 (0.50)</td>
<td>1.02 (0.78)</td>
<td>1.47 (0.30)</td>
</tr>
<tr>
<td>$35,000 - $49,999</td>
<td>1.23 (0.35)</td>
<td>0.94 (0.57)</td>
<td>0.74 (0.77)</td>
<td>1.23 (0.38)</td>
</tr>
<tr>
<td>$50,000 - $74,999</td>
<td>1.83 (0.45)</td>
<td>0.87 (0.57)</td>
<td>1.40 (0.87)</td>
<td>2.50 (0.35)**</td>
</tr>
<tr>
<td>&gt; $75,000</td>
<td>3.49 (0.39)**</td>
<td>1.21 (0.60)</td>
<td>2.12 (0.89)</td>
<td>3.13 (0.38)**</td>
</tr>
<tr>
<td>Occupation (Referent: Employed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployed</td>
<td>0.90 (0.27)</td>
<td>0.79 (0.28)</td>
<td>0.89 (0.43)</td>
<td>0.47 (0.41)</td>
</tr>
<tr>
<td>Disabled</td>
<td>1.11 (0.40)</td>
<td>0.44 (0.37)*</td>
<td>0.72 (0.57)</td>
<td>0.85 (0.41)</td>
</tr>
<tr>
<td>Retired</td>
<td>0.94 (0.23)</td>
<td>0.65 (0.26)</td>
<td>0.69 (0.32)</td>
<td>0.72 (0.19)</td>
</tr>
<tr>
<td>Other</td>
<td>1.51 (0.69)</td>
<td>0.82 (0.57)</td>
<td>1.56 (1.16)</td>
<td>0.36 (0.69)</td>
</tr>
<tr>
<td>Marital Status (Referent: Married)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>0.92 (0.16)</td>
<td>0.72 (0.18)</td>
<td>1.18 (0.23)</td>
<td>1.00 (0.21)</td>
</tr>
<tr>
<td>Most Recent Check-up (Referent: Within the past year)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2 years</td>
<td>0.79 (0.27)</td>
<td>0.80 (0.20)</td>
<td>0.98 (0.32)</td>
<td>0.54 (0.25)*</td>
</tr>
<tr>
<td>3 years or more</td>
<td>0.48 (0.29)**</td>
<td>0.83 (0.24)</td>
<td>0.91 (0.48)</td>
<td>0.45 (0.35)*</td>
</tr>
<tr>
<td>Health Insurance (Referent: Yes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1.02 (0.61)</td>
<td>0.41 (0.40)*</td>
<td>0.54 (0.66)</td>
<td>0.74 (0.51)</td>
</tr>
<tr>
<td>Caregiver (Referent: Yes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>0.87 (0.24)</td>
<td>0.77 (0.20)</td>
<td>0.95 (0.23)</td>
<td>0.71 (0.24)</td>
</tr>
</tbody>
</table>
Advancements in digital technologies offer unique opportunities for patients to remotely monitor their health and to share these data to inform clinical decision-making. Data sharing could also be accelerated by mitigating infrastructure and technical barriers such as access to broadband Internet and providing technical support to patients to assist them with accessing and using technologies that facilitate tracking and data sharing. Efforts could include tailored patient training materials for various communication modalities, such as e-mail, personal health record, or text messaging since a patient’s preference and proficiency for using these may vary based upon sociodemographic attributes such as age, health literacy, or technology usage. Trustworthiness and privacy may also limit patients’ willingness to share
data with health professionals, so transparent policies that are described in patient-friendly language could help improve patients’ understanding and comfort with data sharing.

Additionally, reimbursement and organizational policies could address barriers to access and information sharing. While much progress has been made to facilitate data sharing, continued efforts focusing on reimbursement, interoperability, and addressing health professionals’ concerns about information sharing will help to advance information sharing. Centers for Medicare & Medicaid Services has made progress in reimbursing health care professionals for communicating and monitoring patients from home with the introduction of three new CPT codes in the 2019 Physician Fee Schedule and Quality Payment Program\textsuperscript{28}. These new codes allow for reimbursement for 20 minutes or more of communication with the patient or caregiver per month, remote monitoring of physiological parameters (e.g., weight, blood pressure, pulse oximetry) with daily recordings or programmed alerts, and patient education and set-up of the devices\textsuperscript{28}. This new reimbursement landscape provides opportunities for enhanced patient-health professional communication, but it is vital to address health professional concerns such as effective use and monitoring of these new data streams\textsuperscript{29-31}. For example, providing meaningful visualizations and summarizations of the data as well as incentives to incorporate PGHD into routine clinical workflows could help with the use of these data in clinical care\textsuperscript{31}. Since reimbursement is tied to patient experience through Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) scores\textsuperscript{32}, there is also an increasing focus on improving patient satisfaction. Providing access to health tracking and information sharing modalities based on individual’s health or technology needs could improve the patient experience through enhancing communication and care management. In the current healthcare landscape, it is still challenging to operationalize the collection and exchange of digital data across healthcare systems for clinical care, quality, and research. The Office of the National Coordinator for Health Information Technology released a 2019 Interoperability Standards Advisory that updated standards for health data to encourage data exchange among patients and health professionals\textsuperscript{33}. However, as the U.S. healthcare system continues to shift towards value-based care, greater efforts to increase patient engagement with tracking and data sharing are necessary, and these should align with quality and payer measures.

**Limitations.** The HINTS survey is a cross-sectional survey, which does not allow for examining causal relationships between variables. The survey is self-reported and lacks variables that may be highly correlated with health tracking and information sharing, such as health literacy, patient activation or engagement, or sociocultural factors. The survey does not assess all the potential modalities of health tracking and communication such as non-electronic mediums (i.e., paper-pencil tracking) which could also be useful in managing conditions. Despite these limitations, this study contributes to health tracking and communication research by expanding upon how various communication modalities are associated with sociodemographic characteristics and chronic conditions in a nationally representative sample of U.S. adults.

**Conclusions**

Patient access to technologies that enable tracking and sharing of health information is important to facilitate personalized and collaborative care, particularly for patients with chronic conditions. Sociodemographic differences exist across health tracking behaviors and information sharing modalities. Access to the Internet, electronic devices, or smartphones is limited among patients with chronic conditions. Additionally, access to online medical records is still limited to all patients overall. As consumer health technologies become more sophisticated, additional research is needed to determine whether these further widen health disparities, particularly for patients with chronic conditions. Future research should also examine longitudinal trends in tracking health and information sharing as the context and needs of patients evolves over time, as well as their impact on health outcomes.

**References**


Leveraging Contextual Information in Extracting Long Distance Relations from Clinical Notes

Hong Guan, MS and Murthy Devarakonda, PhD

Biomedical Informatics, College of Health Solutions
Arizona State University, Tempe, AZ

Abstract
Relation extraction from biomedical text is important for clinical decision support applications. In post-marketing pharmacovigilance, for example, Adverse Drug Events (ADE) relate medical problems to the drugs that caused them and were the focus of two recent shared challenges. While good results were reported, there was a room for improvement. Here, we studied two new improved methods for relation extraction: (1) State-of-the-art deep learning contextual representation model called BERT, Bidirectional Encoder Representations from Transformers; (2) Selection of negative training samples based on the “near-miss” hypothesis (the Edge sampling). We used the datasets from MADE and N2C2 Task-2 for performance evaluation. BERT and Edge together improved performance of ADE and Reason (indication) relations extraction by 6.4-6.7 absolute percentage (and error rate reduction of 24%-28%). ADE and Reason relations contained longer text between the entities, which BERT and Edge were able to leverage to achieve the performance improvement. While the performance improvement for medication attribute relations was smaller in absolute percentages, error rate reduction was still considerable.

Background and Significance
Relation extraction is an important problem in biomedical NLP because it helps to identify critical elements of patient care, such as the medications causing adverse drug events, reasons for treatments, and bases for diagnosis, outcomes, and treatment changes, from clinical text. Even in the general domain NLP, relation extraction is an active and ongoing research area. We studied two new strategies for exploiting contextual information to improve relation extraction performance.

One strategy was to use the state-of-the-art neural network model called BERT, Bidirectional Encoder Representations from Transformers, which was shown to achieve significant performance improvement in general domain NLP tasks over task-specific neural architectures that used word2vec, GloVe, or ELMo word embeddings as features. BERT achieves performance improvement in two ways. First, by using multiple layers of bidirectional Transformer encoder blocks which are based on the self-attention model and second by using a learning model that predicts randomly masked tokens in a sequence, called Masked Learning Model (MLM) which is based on the Cloze procedure in Journalism. The self-attention model was shown to better leverage context compared to the neural networks that use LSTMs (Long Short-Term Memory neural networks) and CNNs (Convolutional Neural Networks). Furthermore, BERT pre-trained model only required fine-tuning with task-specific training data and a simple feed forward layer with softmax to predict relation labels.

The second strategy was to use the “near-miss” hypothesis in selecting negative training samples. Typically, negative samples far exceed positive samples in a training set, and the standard approach is to down sample using random selection. But, the strategies used in active learning suggest potential benefits of using near miss samples. A near miss is a negative sample that differs from the learned concept in only a small number of significant points. In psychology of game playing and in AI learning and reasoning, it was observed that the near misses have a significant effect on the outcome. Here, we propose the Edge sampling, which selects negative samples such that relation entities in it are at the “edge” of (or close to) the corresponding positive sample entities in the text. Thus, an Edge negative sample shares a significant text with the corresponding positive sample and yet differs from it by some text (i.e. follows the near-miss hypothesis).

Two recent biomedical NLP challenges focused on Adverse Drug Event (ADE) relation extraction from clinical notes, Medication and Adverse Drug Events from Electronic Health Records 1.0 (MADE) and National NLP Clinical Challenges Task 2 (N2C2). Two obfuscated examples from MADE are shown in Figure 1. While the top performing systems achieved high accuracy (mid to high 0.9 F measures) in extracting medication attribute relations, such as drug-dosage, drug-route, and drug-strength, they substantially under performed on all important ADE and Reason relations.
(indication) relations (mid to high 0.7 F measures). The ADE and Reason relations typically occur over long distances (e.g. up to 501 words apart in MADE) (see Figure 1) and hence there is a challenge and an opportunity to leverage contextual information in these long-distance relations. Here we, therefore, studied the impact of BERT and the Edge sampling on extracting long distance relations using the two datasets.

Example 1:

Example 2:

Mr. [** Name **] comes back for followup examination to pay for peripheral neuropathy. He is approximately 4 weeks status post last cycle of chemotherapy with bendamustine, bortezomib. Early on with this chemotherapy, he had some nausea and vomiting, but has recovered from the same. His other issue of concern at this time and has been some tingling and numbness of his feet and loss of sensation in the tip of his fingers. This probably is related to bortezomib.

Color keys for entities: drug, strength, duration, and ADE
Color keys for relations (indicated by arrows): drug to duration, drug to strength, ADE to drug positive relation, and ADE to Drug negative relation

Figure 1. Relations in the ADE challenges. Example 1 shows medication attribute relations, and Examples 2 shows a positive ADE to drug relation and a negative ADE to drug relation.

Our results showed that BERT alone achieved substantial improvement over the top performing systems for the ADE/Reason relations and the Edge sampling provided additional improvement over BERT with random sampling. Further analysis showed distinct differences in the relation distances between ADE/Reason and medication attributes relations. Thus, this study established new state of the art performance for such long-distance relations in clinical NLP.

Methods

Datasets

The MADE dataset contained 1092 de-identified clinical notes of 21 cancer patients. Each note was annotated with medication information (i.e. drug name, dosage, route, frequency, and duration), ADEs, indications (reasons), other signs and symptoms (SSLIFS), and relations among those entities. The data was split into a training set of 900 notes and a test set of 180 notes.

The N2C2 dataset consisted of 505 discharge summaries from the MIMIC-II clinical care database. Each note was annotated with drug names, dosages, durations, and other entities, and relations of drugs with adverse drug events and other entities. The dataset was split into a training set of 303 notes and a test set of 202 notes.

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Training Dataset</th>
<th>Test Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive Samples</td>
<td>Possible Negative Samples</td>
</tr>
<tr>
<td>ADE-Drug</td>
<td>2,057</td>
<td>27,348</td>
</tr>
<tr>
<td>Reason-Drug</td>
<td>4,530</td>
<td>54,322</td>
</tr>
<tr>
<td>Duration-Drug</td>
<td>903</td>
<td>9,998</td>
</tr>
<tr>
<td>Severity-SSLIF</td>
<td>3,459</td>
<td>25,656</td>
</tr>
<tr>
<td>Dosage-Drug</td>
<td>5,150</td>
<td>45,578</td>
</tr>
<tr>
<td>Frequency-Drug</td>
<td>4,407</td>
<td>39,923</td>
</tr>
<tr>
<td>Route-Drug</td>
<td>2,544</td>
<td>20,776</td>
</tr>
<tr>
<td>Form-Drug</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Strength-Drug</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>23,048</td>
<td>223,715</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Training Dataset</th>
<th>Test Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive Samples</td>
<td>Possible Negative Samples</td>
</tr>
<tr>
<td>ADE-Drug</td>
<td>1,061</td>
<td>4,430</td>
</tr>
<tr>
<td>Reason-Drug</td>
<td>4,991</td>
<td>29,751</td>
</tr>
<tr>
<td>Duration-Drug</td>
<td>642</td>
<td>3,305</td>
</tr>
<tr>
<td>Severity-SSLIF</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Dosage-Drug</td>
<td>4,206</td>
<td>20,581</td>
</tr>
<tr>
<td>Frequency-Drug</td>
<td>6,303</td>
<td>48,993</td>
</tr>
<tr>
<td>Route-Drug</td>
<td>4,995</td>
<td>34,370</td>
</tr>
<tr>
<td>Form-Drug</td>
<td>6,647</td>
<td>34,011</td>
</tr>
<tr>
<td>Strength-Drug</td>
<td>6,703</td>
<td>45,766</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>35,548</td>
<td>221,217</td>
</tr>
</tbody>
</table>

Detailed characteristics of the two datasets are shown in Tables 1 and 2, which include the number of positive relations and potential negative relations in the training and test sets broken down by relation types. The ratios of potential negative relations to positive relations are also shown in the tables. It can be seen that the ratios are high in the training data of both sets, i.e. substantially more potential negative relations than the positive relations, and the ratios are even
higher in the MADE training dataset. Specifically, the ratio is 13.3 for the ADE-Drug relation, 12.0 for Reason-Drug, and 9.7 overall in the MADE training dataset. These high ratios indicate an opportunity to strategically select negative samples to optimize training.

**BERT (Bidirectional Encoder Representations from Transformers)**

BERT is a new exciting development in neural network models research, demonstrating significantly improved state-of-the-art performance on various general domain NLP tasks, including sentence classification which is relevant to us here. BERT is a pre-trained model that produces sequence (e.g., sentence) and word level representations, which can be fine-tuned for task-specific outcomes such as relation classification and concept extraction. Only a simple feed-forward network with a softmax layer is needed to process the BERT output for task-specific objectives.

In the last few years, word2vec has become the de facto standard for producing feature representations of words in biomedical NLP, which are then processed by task-specific neural architectures such as RNNs, LSTMs, CNNs, and heavily-engineered combinations thereof. On occasion, GloVe or ELMo were used instead to generate word representations, but still required complicated task-specific neural network engineering. BERT approach is to do away with the task-specific architectures and provide a broadly applicable pre-trained model which only need to be fine-tuned for the task, using task-specific training data. Other such approaches were also proposed, for example, OpenAI Generative Pre-trained Transformer (GPT), but BERT was shown to outperform them with a multi-layer bidirectional architecture.

As shown in Figure 2a, BERT uses layers of neural network components known as Transformer encoders, shown as $Tm$ in the figure, to generate representations of input sequences in the output. Each BERT layer processes its input sequence in the forward and backward directions simultaneously, using a novel pre-training objective known as the masked learning model (explained later). The BERT Transformer encoder contains two sublayers (see Figure 2b), the first sublayer is a multi-head self-attention mechanism that allows modeling of the context for each word position and the second is a feed-forward network that provides non-linear activation. The encoder architecture also included a residual connection around each sublayer which was shown to simplify optimization and a sublayer of normalization which reduced computational requirements. Fundamentally, the attention function provides a mapping of a query and a key-value pair to an output. Intuitively and as applied here, an attention layer produces output (say, a word representation) that is based on any arbitrary word positions (of the input sequence) by comparing each
sequence member with each other sequence member (self-attention) and producing a series of probability distributions to assign importance. Multi-headed attention can simultaneously optimize for different input combinations.

As mentioned earlier, in pre-training, BERT uses a novel pre-training objective known as the masked learning model (MLM), where some random words in the input are masked, and the pre-training objective is to predict the original word based on the context. In other approaches, typically next word prediction was used, which limited a multi-layered model to process the input either in the forward direction only or process in the forward and backward directions separately and then aggregate the representations – both these approaches fail to leverage the forward and backward contexts at the same time. The use of MLM was a key invention that enabled simultaneous bidirectional input processing in a multi-layer model, without allowing words-to-be-predicted appearing in the input of an upper layer.

The BERT model we used here came pre-trained with BookCorpus and English Wikipedia (general domain corpus). Corpus words were tokenized using the WordPiece dictionary of 30,000 words and as needed words were split into pieces using # (two hash marks). Word piece representations using biomedical corpus were not readily available at the time of our study, however, recently a BERT model was pre-trained on a biomedical corpus and it was shown to improve entity extraction. In the future, we plan to study performance of this BERT model, that was pre-trained on biomedical corpus, in relation extraction.

We employed BERT in its base configuration of 12 layers (~24 sublayers), 768 hidden size, 12 self-attention heads. In our configuration, BERT produced representations for each word (token) in the input as well as a single sequence representation (shown as CLS). This sequence classification representation from the top layer of BERT (see Figure 2a) was used as the input to a feed-forward layer containing one hidden layer. A softmax layer provided the final relation classification label for the pair of entities in the input sequence.

In our method, as shown in Figure 2c, the input to BERT was a sequence of words that started with the first entity of a relation and ended with the second entity of the relation. As previous studies demonstrated, this is a convenient choice since sentence segmentation of clinical notes text is error prone due to embedded lists and tables which are not well handled by the standard NLP code such as NLTK. We were also constrained by our BERT model input limit which was 512 word-pieces. Entity spans were further marked using the entity tags.

In the training (fine-tuning) stage, the input sequences included all positive samples in the gold standard and an equal number of negative samples, that were either randomly down sampled from all possible negative samples or Edge sampled as described in the next section. Only the sequences that were less than or equal to a heuristically determined (in the validation stage) maximum sequence length were used in training.

The Edge Sampling

In the MADE training dataset, there were 13.3 times as many potential negative samples as the positive samples for the ADE relation. In the N2C2 training dataset the ratio was 4.2. Ideally, the negative and positive samples should be equal for optimum model training. The new sampling method we propose here, called the Edge sampling, tries to select negative samples that share the largest common text (and hence context) with the corresponding positive samples, i.e. near-miss samples.

The Edge sampling works as follows. For each entity of a gold standard relation, the immediately preceding and following entities of the same type as the gold standard entity are identified (if they exist) in the text. From these four neighboring entities and from the original two entities from the gold standard, valid relations are formed by having one (and exactly one) of the gold standard entities as a part of the relations. Valid means the relation meets its entity type requirements and the entities occur in a sequence of words in the document. If entities in a gold standard relation are of different types (as in the case of relations in MADE and N2C2), at most four such new relations can be formed. This is illustrated in Figure 3. Note that, depending on the location and entities in a clinical note, we may find four or less, including zero, such new relations. If any of these relations were already in the gold standard or were picked already as a negative sample, they were removed from the list. Among the remaining relations, the Edge sampling randomly picks one.

If an Edge sample cannot be found for a positive sample, a random negative sample (that was not already picked) was selected from the rest of possible negative samples. Negative example selection takes place only in the training phase. During the evaluation and validation phases, all potential relations were assessed by the model and outcome measured.

Due to practical limitations, our model enforces a maximum sequence length, which is heuristically determined during the validation phase. In training, positive samples longer than the max length or ignored, and similarly in the Edge
Entity types of a relation type R are: $T_1$ and $T_2$

| Entity instances of type $T_1$ in the text (p and f indicate immediately preceding and following i) | $T_{1p}^i$, $T_{1i}^f$ |
| Entity instances of type $T_2$ in the text (p and f indicate immediately preceding and following i) | $T_{2p}^i$, $T_{2i}^f$ |
| Gold standard positive relation of type $R$, $R_i$ | $[T_{i1}, T_{2i}]$ |
| For $R_i$, Edge relation sequences are: $R_1$, $R_2$, $R_3$, $R_4$ | $[T_{11}, T_{21}], [T_{11}, T_{2f}], [T_{1p}, T_{2i}], [T_{1f}, T_{2i}]$ |

Figure 3. The Edge sampling illustrated.

We fine-tuned our BERT-based neural network model separately for MADE and N2C2, each with and without the Edge sampling. We therefore trained (fine-tuned) the model four different times and tested each one. We mostly adopted the default settings of BERT hyperparameters – i.e., training batch size of 16, 10 epochs, and a learning rate of $2e-5$. However, we experimentally determined the optimum sequence length using validation sets - 20% of clinical documents from the N2C2 training set and 10% of clinical documents from the MADE training set. The fine-tuned models were tested on the full MADE and N2C2 test datasets.

For each of the four experiments, we calculated standard Recall, Precision, and F measures individually for all relations as well as for all the relations combined. We compared our results with the published results of the systems that performed best in the task of relation extraction given gold entity labels from the two challenges: the University of Utah system$^{23}$ for MADE (denoted as the MADE_Best), and the UTHealth developed system$^{24,25}$ for N2C2 (denoted as the N2C2_Best). Since the published results did not aggregate performance for the ADE and Reason relations, we obtained their weighted average, weighted by the number of relations evaluated in the test phase, from individual relation results. We used two metrics for comparison: (1) Absolute F measure difference; and (2) Error rate reduction in the F measure achieved by model Y compared to model X, which is calculated as:

$$\text{error rate reduction} = \frac{F_Y - F_X}{1 - F_X}$$

where $F_Y$ and $F_X$ are the F measures of models Y and X respectively. While the absolute F measure difference shows the net improvement in the measure, the error rate reduction is a sound relative measure that shows reduction in the remaining performance gap of the previous model. It is usually expressed as a percentage. Recent studies in the general domain NLP have adapted this metric for effective comparison.$^{1,2}$

We suspected that for certain relation types, such as the medication attributes, the distance between the entities may be short and in which case the potential to leverage contextual information is rather limited. The two datasets might also be different in terms of the distances between entities, since N2C2 contains discharge summaries whereas MADE contains clinical notes. In order to quantify such differences, we studied the sequence length distributions of positive samples, Edge samples, and all negative samples for each dataset and plotted their cumulative distribution frequency.

### Results

Table 3 shows performance evaluation on the MADE dataset. Precision, recall, and F measures were shown for BERT (using random negative instances sampling) and BERT with the Edge sampling (shown as BERT+Edge) for each relation type, for the aggregate of ADE and Reason relations, and for all relations together. The table also shows...
absolute F measure differences between MADE_Best and BERT, BERT and BERT+Edge, and between MADE_Best and BERT+Edge. Percentage error rate reduction was shown between MADE_Best and BERT+Edge.

In terms of absolute F measure, BERT achieves 1.1% improvement overall compared to MADE_Best, but 4.1% improvement for the ADE and Reason relations. Edge improved the overall F measure by additional 1.6% compared to MADE_Best and by additional 1.9% for the ADE and Reason relations. BERT improved the Reason relation extraction F measure significantly (by 6.6%), while Edge improved the ADE relation F measure significantly (by 3.3%). Performance improvement from BERT+Edge over MADE_Best is substantial: absolute F measure improved by 2.7% overall and by 6.4% for ADE and Reason. Notice that the error rate was reduced by 22.7% overall and 23.6% for ADE and Reason. The error rate reduction was substantial for most relations.

**Table 3. Performance results for the MADE dataset**

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>MADE_Best (P/R/F)</th>
<th>BERT (P/R/F)</th>
<th>BERT over MADE_Best: Δ F measure</th>
<th>BERT+Edge (P/R/F)</th>
<th>BERT+Edge over BERT: Δ F measure</th>
<th>BERT+Edge over MADE_Best: Δ F measure</th>
<th>err. reduced%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADE-Drug</td>
<td>0.787/0.683/0.731</td>
<td>0.652/0.848/0.737</td>
<td>0.006</td>
<td>0.730/0.814/0.770</td>
<td>0.033</td>
<td>0.039</td>
<td>14.5%</td>
</tr>
<tr>
<td>Reason-Drug</td>
<td>0.780/0.739/0.758</td>
<td>0.728/0.948/0.824</td>
<td>0.066</td>
<td>0.772/0.904/0.833</td>
<td>0.009</td>
<td>0.075</td>
<td>31.0%</td>
</tr>
<tr>
<td>Duration-Drug</td>
<td>0.937/0.912/0.924</td>
<td>0.759/0.973/0.853</td>
<td>(0.071)</td>
<td>0.946/0.952/0.949</td>
<td>0.096</td>
<td>0.025</td>
<td>32.9%</td>
</tr>
<tr>
<td>Severity-SSLIF</td>
<td>0.911/0.962/0.936</td>
<td>0.959/0.971/0.965</td>
<td>0.029</td>
<td>0.966/0.977/0.971</td>
<td>0.006</td>
<td>0.035</td>
<td>54.7%</td>
</tr>
<tr>
<td>Dosage-Drug</td>
<td>0.957/0.962/0.960</td>
<td>0.924/0.983/0.952</td>
<td>(0.008)</td>
<td>0.939/0.970/0.954</td>
<td>0.002</td>
<td>(0.006)</td>
<td>(15.0%)</td>
</tr>
<tr>
<td>Frequency-Drug</td>
<td>0.971/0.923/0.947</td>
<td>0.935/0.966/0.950</td>
<td>0.003</td>
<td>0.942/0.967/0.955</td>
<td>0.005</td>
<td>0.008</td>
<td>15.1%</td>
</tr>
<tr>
<td>Route-Drug</td>
<td>0.961/0.921/0.941</td>
<td>0.941/0.937/0.958</td>
<td>0.017</td>
<td>0.957/0.980/0.968</td>
<td>0.01</td>
<td>0.027</td>
<td>45.8%</td>
</tr>
<tr>
<td>Form-Drug</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strength-Drug</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.903/0.859/0.881</td>
<td>0.839/0.953/0.892</td>
<td>0.011</td>
<td>0.880/0.938/0.908</td>
<td>0.016</td>
<td>0.027</td>
<td>22.7%</td>
</tr>
<tr>
<td>ADE+Reason</td>
<td>0.783/0.712/0.746</td>
<td>0.700/0.911/0.791</td>
<td>0.041</td>
<td>0.757/0.871/0.810</td>
<td>0.019</td>
<td>0.064</td>
<td>23.6%</td>
</tr>
</tbody>
</table>

Table 4 shows performance evaluation on the N2C2 dataset in the same way as in Table 3. The general trend of the results is similar to that of the MADE results but notably the absolute F measure improvement for the aggregate of all relations is somewhat smaller. BERT achieves a modest 0.7% F measure improvement overall and Edge managed to add additional 0.4% F measure to a total of 1.1% improvement. On the other hand, BERT improved the F measure of the Reason relation significantly, by 6.8%. The highest Edge F measure improvement was for the ADE relation, adding an additional 1.3% to BERT. While the F measure improvement from BERT+Edge over MADE_Best for all relations combined is relatively small, 1.1%, but the error rate reduction was a substantial 18.3% indicating that the
improvement in relative terms is still significant. The F measure improvement for the aggregate of ADE and Reason relations is 6.7%, and the corresponding error rate reduction was 28.3%, both of them indicate significant improvement. As in the case of MADE results, the error rate reduction was substantial for most relations. Precision and recall details were publicly unavailable for the N2C2_Best at the time of writing this paper.

**Statistical significance test:** Previous studies have used the McNamara test (and is generally accepted as a good test) for determining the statistical significance of F measure improvement of an NLP task. The test requires the contingency (confusion) table from the performance study. Using the data in our study, we determined that the F measure improvement with the Edge sampling was statistically significant at p<0.001 for the MADE dataset for all relations combined and for ADE+Reason relations. For the N2C2 dataset, the improvement was also significant at p<0.001 for all relations combined but was only significant at p<0.03 for ADE+Reason relations. We could not determine statistical significance of performance improvements relative to the MADE_Best and N2C2_Best models because the contingency tables for them are not publicly available at the time of this article.

Another important observation from Tables 3 and 4 is that the Edge sampling consistently improved precision, while often losing ground on recall. Edge improved precision for the overall and ADE+Reason by 4.1% and 5.7% for the MADE dataset, and by 1.2% and 2.3% for the N2C2 dataset respectively. Recall reduced by small percentages across the broad. These results indicate an important characteristic of our Edge sampling approach. We also note that both BERT and BERT+Edge consistently improved recall over MADE_Best.

The sequence length distributions of the positive, all negative, and Edge-sampled relations in the training datasets of MADE and N2C2 are shown in Figures 4 and 5 respectively. We showed the distributions for the ADE and Reason relations and for all the rest of relations. We showed detailed statistics of the distributions in Table 4. Two important observations can be made from the Figure and the Table.

First, the ADE and Reason relation lengths are significantly different from the lengths of the rest of the relations, especially those of the potential negative samples. In the MADE dataset, the median length of the ADE and Reason negative relations is 155 words, which is 4.3 times the median length of the rest of the relations (see Table 5). In N2C2, the ratio is more modest 1.7 times but it is still significant.

Second, the negative sample relation lengths in MADE are significantly longer than in N2C2 for the ADE and Reason relations. The median length of the relations in MADE is 3.04 times the median length of N2C2 negative relations, i.e. 155 versus 51. The median lengths of the rest of the relations are similar in both datasets.
While some differences exist between the two datasets, the ADE and Reason relations are consistently longer than the rest, and therefore, these sequences are likely to contain more context, which can be leveraged by BERT and the Edge sampling. The Figure and the Table also show that the Edge sampling, reduces the negative sample lengths. For example, the median lengths for ADE/Reason are 55 and 32 for the two datasets, which are substantially smaller than the median lengths of all negative samples (i.e. 155 and 51).

Discussion

The methods we used here, BERT and the Edge sampling, better leverage contextual information to improve relation extraction compared to the top performing systems from the shared challenges, MADE and N2C2. BERT makes better use of context in two ways:

1. Multiple layers of Transformer encoders based on the Attention model, rather than the RNN, LSTM, CNN, or combination models thereof, was shown to better leverage context;

2. Using the Masked Learning Model that randomly masks a word in a sequence for conditioning word representations to simultaneously analyze the input bidirectionally, rather than predicting merely an association of words (as in word2vec), or the next words (as in OpenAI GPT), or even a concatenation of representations predicting next and preceding words (as in ELMo).

In addition, BERT only required fine-tuning of a pre-trained model rather than complex task-specific neural networks that use word representations as features thus achieving transfer learning from a large text corpus. For these reasons, BERT in our study performed better than the top performing systems from the challenges. It is interesting to note that the MADE_Best employed a carefully feature-engineered Random Forest and N2C2_Best used complex composition of neural networks consisting of LSTMs and CNNs, and BERT improved upon both approaches.

Selection of negative samples in model training is known to be a challenge, especially when relation entities are multiple sentences apart which gives rise to a very large number of potential negative samples. Most previous studies simply down sampled the larger population, but one study considered a feature-engineered Alternating Decision Tree machine learning model for selecting candidate samples, both for training and testing. The Edge sampling, by preferring negative training samples that share significant context with positive samples, provides a simpler and in combination with BERT higher accuracy compared to the previous candidate selection approach on the MADE dataset.

In can be observed from the results that the performance improvement achievable from contextual information depends on the length of context between relation entities. Quantitatively, the relation distance differences between ADE/Reason and medication attribute relations can be seen in Figures 4 and 5, and in Table 5. Qualitatively, as can be seen from the examples in Figure 1, medication attributes tend to appear in short segments and often have a simple and easy to recognize patterns. These simple patterns can be easily recognized more easily and indeed, the previous studies have achieved very high F measures, often above 0.95, by recognizing such patterns. Whereas, the entities of ADE and Reason appear anywhere and in complex discourse in a clinical document. These typically longer relations offer an opportunity for the sophisticated neural architecture of BERT to create better representations, and hence increase F measure accuracy substantially. It should however be noted that the error rate reduction is significant across most relations with BERT plus Edge.

The Figures 4 and 5, and the Table 5 also show that the Edge sampling, not only takes advantage of the contextual information (by definition) but also reduce the negative sample lengths and make their lengths similar to the positive samples. As we noted earlier, the Edge sampling disproportionately improves precision over recall, which we plan to study further to understand the reasons for it and how it can be leveraged further.

Conclusion

This study makes the following important contributions:

1. We showed that the BERT model, which leverages bidirectional contextual information with multi-layer Transformers, requiring only fine-tuning can provide excellent performance in biomedical relation extraction without complicated, task-specific neural network designs containing RNNs, LSTMs, and CNNs.

2. We showed that the “near-miss” based the Edge sampling of negative instances, rather than random selection, can improve training and therefore model performance especially when there is a large population of potential negative samples to choose from and when relation sequences are long enough to form “near-misses” from positive samples.
3. Both BERT and Edge use contextual information in long distance relations to achieve significant performance improvement. The performance improvement, compared to the previous methods, was substantial for such relations (i.e. ADE and Reason relation): 6.4% absolute F measure improvement (23.6% error reduction) for the MADE dataset and 6.7% absolute F measure improvement (28.3% error reduction) for the N2C2 dataset.

Conflict of Interest
The authors do not have any conflicts of interest.

References
Timing Considerations for Noninvasive Vagal Nerve Stimulation in Clinical Studies

Nil Z. Gurel*, MS¹, Asim H. Gazi*, BS¹, Kristine L. Scott, BS¹, Matthew T. Witthodt, PhD², Amit J. Shah, MD, MPH³,⁴,⁵ Viola Vaccarino, MD, PhD³,⁴, J. Douglas Bremner, MD²,⁵,⁶, Omer T. Inan, PhD¹,⁷

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA; ²Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA; ³Department of Epidemiology, Rollins School of Public Health, Atlanta, GA; ⁴Department of Internal Medicine, Emory University School of Medicine, Atlanta, GA; ⁵Atlanta VA Medical Center, Decatur, GA; ⁶Department of Radiology, Emory University School of Medicine, Atlanta, GA; ⁷Coulter Department of Bioengineering, Georgia Institute of Technology, Atlanta, GA

Abstract

Noninvasive vagal nerve stimulation (n-VNS) devices have the potential for widespread applicability in improving the well-being of patients with stress-related psychiatric disorders. n-VNS devices are known to affect physiological signals, and, recently, they have been employed in various protocols involving both acute and longitudinal applications. However, questions regarding response time, “dosage,” or optimal treatment paradigms remain open. Prior work evaluated noninvasively obtained biomarkers that quantify the stimulation efficacy based on the changes in autonomic tone in a randomized double-blind study. In this work, we extend the state-of-the-art by investigating the onset of action for n-VNS in these same physiological biomarkers through a three-day clinical trial, including 233 administrations on 24 human participants, with and without immediately preceding acute traumatic stress. Determining n-VNS latency serves as a substantial step toward optimizing stimulation delivery with higher temporal resolution for personalized neuromodulation.

Introduction

The vagus nerve is a major component of the autonomic nervous system, consisting of afferent (sensory) and efferent (motor) connections to regulate autonomic tone and maintain homeostasis¹. Electrical stimulation of the vagus nerve is hypothesized as a “bottom-up” therapy that acts on the vagal afferents through the nucleus tractus solitarius (NTS), which projects to the brain areas that regulate mood and emotion, such as the amygdala, hippocampus and prefrontal cortex¹,³ (see Figure 1 for the representation of a simplified mechanism of action). Through the modulation of vagal afferents, efferent connections carry information to regulatory organs that modulate autonomic state, such as the heart. Surgical implants for vagal nerve stimulation (VNS) appear effective in the treatment of several psychiatric and cardiovascular disorders, such as epilepsy, depression, migraines, and heart failure⁴–¹². Though high variability in treatment response has been observed with open-loop approaches, various closed-loop algorithms have been proposed for implantable VNS devices using measures related to the heart rate¹³–¹⁸ or neural response¹⁹. Indeed, compared to their open-loop counterparts, the closed-loop implementations of these algorithms on VNS implants in the market have proven effective in improving clinical outcomes²⁰. However, widespread adoption is limited due to the potential for adverse events during implantation and the normal cost and practical challenges associated with implantable devices²¹,²².

Noninvasive VNS (n-VNS) devices provide advantages in terms of usability and low-cost, with proven effects on central brain regions²³–²⁶, cytokines, and patient physiology²⁷–²⁹. Therefore, n-VNS devices have the potential to address many shortcomings of their implantable counterparts, potentially improving patient adherence and broadening use³⁰. Their operation, however, is currently limited to open-loop applications prescribed by the medical professional. Translating implantable closed-loop technologies to the noninvasive neuromodulation realm is not straightforward; the physiological measurements should also be obtained noninvasively, leveraging wearable sensing modalities. Moreover, the effects of n-VNS on human physiology would likely differ from the transient responses to implantable

*Authors equally contributed to this work.
Figure 1. Simplified representation of n-VNS mechanism of action. The understanding of n-VNS kinetics on noninvasively obtained physiological parameters may enable optimization of n-VNS delivery in unsupervised settings. NTS: nucleus tractus solitarius.

VNS observed and quantified in preclinical studies due to the skin-electrode impedance, the filtering effects of the skin as a barrier between the stimulation site and the vagus nerve, and the time it takes to adjust the stimulation amplitude per subjective tolerance[^31]. For instance, implantable VNS has sub-second effects in anesthetized animals[^32],[^33] and has been modeled for electrodes that are in direct contact with the nerve[^34],[^35]. Latency for the physiological effects of n-VNS would likely be greater than implantable stimulation due to the mentioned differences regarding the electrode-skin-tissue barrier. Determining the onset of action for n-VNS provides pertinent information on n-VNS kinetics in continuous physiological signals germane to design specifications for wearable sensing systems; this, in turn, advances us one step further toward closing the loop for improved n-VNS treatment quality.

Prior work introduced and evaluated downstream cardiovascular and peripheral biomarkers of cervical n-VNS that quantify the physiological response to acute treatment using noninvasively obtained signals; the effects of n-VNS when applied after laboratory-induced psychological stress were also explored[^36]-[^38]. Statistics involving said biomarkers were presented as static outcomes based on average changes from a baseline state upon acute stress application. These biomarkers included heart rate (HR) as a measure of both sympathetic and parasympathetic influences, the pre-ejection period (PEP) of the heart as a measure of cardiac contractility and sympathetic activity, and the amplitude of peripherally measured photoplethysmogram (PPG) signals as a peripheral sympathetic measure. In this work, we extend the state-of-the-art by making the following contributions: (i) We analyze when these biomarkers change during stimulation on 24 human participants in a double-blind study, half undergoing active n-VNS and half undergoing sham stimulation; (ii) We propose a method to identify the latency between n-VNS initiation and the onset of n-VNS related physiological changes determined by the outset of expected changes in the instantaneous biomarkers; and (iii) To differentiate the physiological effects of n-VNS from possible placebo effects or other means of modulation, we applied the same latency identification method on the data obtained from the half who received sham stimulus at the same treatment area on the neck; we then compared both groups’ results. Our findings from 233 administrations suggest pertinent timing considerations germane to the design of effective clinical studies involving physiological effects of n-VNS.

Methods

Human Subjects Experiments

Figure 2. Protocol diagram. The protocol included three days: the first day included six traumatic stress prompts followed by immediate n-VNS or sham stimulation and two stimulation administrations without stress. Each of the second and third days included one stimulation administration without stress.
The study examines the physiological effects of n-VNS following acute traumatic stress. Investigating acute traumatic stress has clinical importance, as exposure to traumatic events can cause strongly encoded intrusive memories, or in contrast, impairments in memory function. These intrusive thoughts may persist in vulnerable individuals, possibly leading to posttraumatic stress disorder (PTSD). As traumatic stress is a trigger for PTSD symptoms, the protocol was designed around laboratory-induced traumatic stress exposure followed by stimulation. Note that stimulation also occurred on its own to understand possible physiological differences. The study was approved by the Institutional Review Boards of Georgia Institute of Technology, Emory University School of Medicine, SPAWAR Systems Center Pacific, and the Department of Navy Human Research Protection Program. A total of 24 adults who have experienced prior psychological trauma (ages mean ± SD: 31 ± 9 years, 12 females) were recruited, and written, informed consent was obtained. Each participant was asked to write his / her personal traumatic memories; later, voice recordings based on these memories were prepared for delivery as acute traumatic stress during the protocol. The traumatic stressors (each lasting for approximately one minute) were delivered audibly through headphones. Upon randomization for double-blind protocol, each participant was assigned to an active n-VNS or sham device having identical appearance, placement, and operation (GammaCore, ElectroCore). Participants, clinical staff, and researchers were blinded to the devices. Active n-VNS devices produce a 25 Hz voltage signal with 5kHz sine wave bursts. Sham devices produce a slow AC biphasic voltage signal with 0.2 Hz square pulses. The stimulation intensity ranges from 0 to 5 arbitrary units (AU), with a corresponding peak output ranging from 0 to 30V for active n-VNS, and from 0 to 14 V for the sham device. During each application, intensity was increased to a tolerable level determined by the participant. The amplitude levels participants received were 3 AU (± 0.8 SD) for active n-VNS and 4.5 AU (± 1 SD) for sham stimulus.

The protocol spanned three days for each participant (see Figure 2 for the protocol summary). On the first day, participants listened to their traumatic stress scripts as acute traumatic stressors (a total of six scripts per participant), and stimulation was applied immediately after each script. To understand the physiological changes for the stress-free (baseline) condition, this day also included two separate n-VNS/sham applications without traumatic stress scripts read immediately prior to stimulation. On both the second and third days, participants received one stimulation without a preceding stress script. Overall, the participants received six stimulations following six traumatic stress prompts and four stimulations without stress. Five participants did not complete all the traumatic stress prompts in the protocol. Excluding the missing data from these participants, there was a total of 137 traumatic stress prompts, followed by 137 stimulation administrations (72 active, 65 sham), and 96 administrations without stress (48 active, 48 sham). Each participant received approximately 10 administrations over the protocol days (233 administrations in total, 120 active, 113 sham).

![Figure 3. Signal processing and feature extraction steps. BPF: bandpass filter; exp. MA: exponential moving average; EA: ensemble average; AO: aortic opening point; PEP: pre-ejection period; HR: heart rate.](image)

**Physiological Sensing**

Noninvasive cardiovascular and peripheral signals were concurrently collected during the protocol. To measure the electrical activity of the heart, three-lead electrocardiography (ECG) signals were collected. As a blood volume pulse measure, finger-based transmissive photoplethysmography (PPG) signals were collected. Chest-wall vibrations (indicative of the mechanical activity of the heart) were measured using seismocardiogram (SCG). ECG and PPG signals were acquired using wireless amplifiers (Bionomadix RSPEC-R and PPGED-R, Biopac Systems). SCG signals were collected using a low-noise accelerometer placed on the mid-sternum (345A32, PCB Electronics, Depew, NY). All data were transmitted to a 16-bit data acquisition system (MP150, Biopac Systems) at a 2kHz sampling rate.

**Signal Processing and Feature Extraction**

Figure 3 summarizes the signal processing and feature extraction steps completed in MATLAB (R2017b, Natick, MA), also detailed in 36. All signals were first bandpass filtered with finite impulse response filters for noise reduction
using 0.6-25Hz, 0.6-40Hz, and 0.4-8Hz frequency ranges for SCG, ECG, and PPG signals, respectively, to maintain consistency with prior literature. The continuous physiological features extracted were HR, PEP, and PPG amplitude. First, R-peaks of ECG signals were located using thresholding. The R-R intervals were then calculated and converted to instantaneous HR in beats per minute (bpm). Second, SCG and PPG signals were segmented (ensemble averaged) referenced to the R-peaks of ECG signals. Third, SCG beats were exponentially moving averaged with a time constant of three beats to reduce the effect of motion artifacts. After segmentation, fiducial points of the beats were located by i) labeling the second peak of SCG beats as aortic opening (AO) point and ii) by finding the global maximum and minimum of PPG beats to extract PPG amplitude. The time interval from ECG R-peak to SCG AO point was extracted as instantaneous PEP (in milliseconds, ms), and the amplitude difference between global maximum and global minimum of each PPG beat was extracted as instantaneous PPG amplitude (in AU).

As the instantaneous variables are prone to corruption by motion artifacts, arrhythmias, and other complex mechanisms, they should not directly be used, as suggested in multiple works examining closed-loop physiological control. Therefore, we smoothed the instantaneous signals using local regression, assigning lower weight to the outliers and using a span of 5% of the total number of data points in a 25-minute interval. Smoothing provides new variables that reflect the transient changes associated with n-VNS.

Figure 4. Annotation diagram. The smoothed instantaneous biomarkers (HR, PEP, PPG amplitude) were plotted from pre-stimulus to post-stimulus. If at least two of the three mentioned changes in the biomarkers occurred, the onset time was marked at the onset of the second change. If no eligible change was observed, the annotation was marked as “absent.”

**Onset of Action Annotation**

The onsets of action following device initiation were annotated manually using the smoothed variables. Manual annotation was completed by three researchers. The researchers were instructed to peruse all participants' data prior to manual labeling. Based on our previous investigations and related implantable VNS literature, the anticipated changes following stimulation compared to the period before stimulation were: i) decrease in HR (i.e., decrease in sympathetic tone or increase in parasympathetic tone), ii) increase in PEP (i.e., decrease in cardiac contractility and cardiac sympathetic activity), and iii) increase in PPG amplitude (i.e., decrease in peripheral sympathetic activity). As multiple sensing modalities were used, occasional noisy measurements existed. Therefore, observing the occurrence of all three of these changes following each n-VNS administration was not expected. Thus, as indicated in in Figure 4, we located the first datapoint that satisfied two of the three aforementioned criteria and marked this time point as the onset time. These criteria were established and agreed upon prior to any manual labeling to ensure guidelines were not contorted to match desired results *a posteriori*. To find the corresponding onset of action, we subtracted the n-VNS start time from the labeled onset time. If only one of the three signals experienced noticeable change, the onset of action was marked as “absent.” Likewise, if no changes at all were observed or if all changes were relatively insignificant compared to noise and normal variation, the corresponding onset of action was marked as “absent.” To allow for independent sampling, we averaged the counts of absent onsets for each participant. Additionally, all annotations were independently performed by the three researchers to later assess the inter-annotator agreements for validation purposes. These were calculated as follows: pairwise agreements between the annotators were calculated in seconds (absolute difference between each onset time annotation) and in counts (agreement percentage for absent
onsets). The final agreement results reported are the overall average absolute difference in annotated onset times and the average percent agreement in absent onsets.

Statistical Analysis
To understand whether baseline characteristics and baseline biomarker measures were comparable between the active n-VNS and sham groups, demographic characteristics (age, gender, weight, height, body mass index) and baseline measures of the biomarkers (HR, PEP, PPG amplitude) were compared using two-sample t-tests for normally distributed continuous variables, Wilcoxon rank-sum tests for non-normal continuous variables, and chi-squared tests for categorical variables. The Shapiro-Wilk test was used to assess normality. Next, the Wilcoxon rank-sum test was used for absent onset count comparison between the groups. A two-sided threshold of p<0.05 was used to indicate statistical significance. All statistical analyses were performed using MATLAB Statistics and Machine Learning Toolbox.

Table 1. Participant demographics and baseline biomarker values in each device group. P: p-value for the comparison of participants’ characteristics between groups. Values represent mean ± SD. BMI: body-mass index. F: female

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Active</th>
<th>Sham</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age [years]</td>
<td>29 ± 7</td>
<td>32 ± 11</td>
<td>0.23</td>
</tr>
<tr>
<td>Sex [F, %]</td>
<td>5F, 41.6%</td>
<td>7F, 58.3%</td>
<td>0.41</td>
</tr>
<tr>
<td>Weight [kg]</td>
<td>77 ± 14</td>
<td>79 ± 13</td>
<td>0.71</td>
</tr>
<tr>
<td>Height [cm]</td>
<td>175 ± 11</td>
<td>172 ± 6</td>
<td>0.49</td>
</tr>
<tr>
<td>BMI [kg / m²]</td>
<td>25 ± 3</td>
<td>27 ± 5</td>
<td>0.34</td>
</tr>
<tr>
<td>HR [bpm]</td>
<td>67 ± 14</td>
<td>61 ± 9</td>
<td>0.22</td>
</tr>
<tr>
<td>PEP [ms]</td>
<td>67 ± 18</td>
<td>75 ± 37</td>
<td>0.48</td>
</tr>
<tr>
<td>PPG Amplitude [V]</td>
<td>0.3 ± 0.3</td>
<td>0.2 ± 0.1</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Findings
Table 1 lists the baseline characteristics of the device groups and associated p-values, proving there are no statistical differences in either demographics or baseline biomarkers. The continuous physiological parameters from two representative participants in the presence of traumatic stress are shown in Figure 5, one participant undergoing sham and the other undergoing active n-VNS. The pre-stimulus values for each of the biomarkers are shown with dashed lines to provide a reference for the predicted deviations.

Table 2 lists the onset of action for active n-VNS for both with and without traumatic stress, as well as the absent onset counts comparing the active and sham device groups. Based on the physiological biomarkers used, the effects of n-VNS were observed 18 ± 7 seconds from the start of n-VNS without stress. When n-VNS was applied after traumatic stress (six traumatic stress scripts followed by six n-VNS per participant, n=12 participants), effects were observed on the biomarkers in a similar latency, resulting in 16 ± 9 seconds. As for the absent onset counts (n-VNS administrations that had not met the criteria), a significant difference exists between the device groups: there were 5 ± 2 absent onsets per participant for the sham group, significantly higher than the active group’s 2 ± 2 absent onset counts (p=0.006). There were 24 and 65 absent onset counts for 120 active n-VNS and 113 sham administrations, respectively. The overall average of inter-annotator agreements resulted in a 4 ± 1 seconds difference between the labeled onset times and a 90 ± 5.5% agreement in absent onset counts.
Figure 5. Continuous physiological parameters showing n-VNS without traumatic stress, for one participant undergoing sham (left) and one participant undergoing active n-VNS stimulus (right). Markers represent the extracted data, lines represent the smoothed data. Shaded regions represent stimulus delivery. Dashed lines show the pre-stimulus averages of the measures. The onset time is marked as soon as two of the three annotation criteria mentioned in Figure 4 are detected.

Table 2. Onset of action and absent onset counts. Values represent mean ± SD.

<table>
<thead>
<tr>
<th>n-VNS administrations</th>
<th>Onset of Action [seconds]</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-VNS with traumatic stress (n=72 administrations)</td>
<td>16 ± 9</td>
</tr>
<tr>
<td>n-VNS without stress (n=48 administrations)</td>
<td>18 ± 7</td>
</tr>
<tr>
<td><strong>Device groups (p=0.006)</strong></td>
<td><strong>Absent Onsets [counts per subject]</strong></td>
</tr>
<tr>
<td>Active n-VNS (n=12 subjects)</td>
<td>2 ± 2</td>
</tr>
<tr>
<td>Sham (n=12 subjects)</td>
<td>5 ± 2</td>
</tr>
</tbody>
</table>

Discussion

This work investigated the onset of action for n-VNS as observed by noninvasive physiological signals. The time durations reported herein serve to estimate the onset of action for n-VNS, which can be used for clinical study design and to provide engineering considerations of closed-loop n-VNS therapy. The continuous signals from representative subjects in Figure 5 show the broad expected changes in the entire sample. The participant undergoing sham stimulus did not experience notable change in HR, PEP, and PPG amplitude during the stimulus, compared to the period before the stimulus. In contrast to the sham participant, the active participant experienced notable physiological modulation during the stimulus: a sudden drop in HR, keeping it below the pre-stimulus range during the stimulation. Additionally, PEP first recovered to the pre-stimulus range during the stimulation, then it exceeded this range by the end of the stimulation, indicating decrease in cardiac contractility and cardiac sympathetic activity. A latent increase in PPG amplitude compared to pre-stimulus was also observed, indicating decrease in peripheral sympathetic activity. These transient changes (decrease in HR, increase in PEP and PPG amplitude) could be interpreted as decreased sympathetic tone during the stimulation. Note that in this active participant, the PPG amplitude remained at a relatively increased level after the stimulus was removed, indicating a sustained decrease in peripheral sympathetic activity even after application. This finding was quite typical in active participants, supporting post-stimulus findings from our previous static analysis study. Another noteworthy set of differences between the active and sham responses deals with changes in these biomarkers prior to stimulus delivery, which can be seen in Figure 5. In particular, we observe a pre-stimulus spike in HR and decrease in PEP. These momentary increases in sympathetic tone prior to stimulus delivery
for active n-VNS could be related to anxiety and was quite typical in the overall sample set. During application, stimulation intensity is increased to as high a level as the subject can tolerate without pain. The active device delivers higher power than the sham device for the same intensity level, which might be perceived as a higher perturbation to the physiology.

As the vagus nerve is a complex structure that affects both sympathetic and parasympathetic activity, the timing of n-VNS effects will vary depending on the type of autonomic response. We observe that the effects in sympathetic measures, PEP and PPG amplitude, for the representative active subject are more latent than HR, another typical occurrence in the overall sample. Also notice that the peak effects in these biomarkers are observed towards the end of the stimulus period. In contrast, the physiologic response due to parasympathetic activation as seen in HR would occur sooner, as it is influenced by both sympathetic and parasympathetic changes. Sympatho-inhibitory phenomena (as observed in PEP and PPG amplitude in the context of this study) are regarded as slower than parasympathetic activation (seen in HR), which goes along with our observations.

**Timing Considerations for n-VNS Studies**

Mental stress studies are particularly important for psychiatry and the cardiovascular disease domain. Physiological responsivity to mental stress has been shown to carry cues on the identity and severity of several conditions such as PTSD, depression, anxiety, and coronary artery disease. Therefore, various types of stressors serve as clinical tools that could be induced in the laboratory under controlled conditions for diagnosis and prognosis purposes. Pairing stressors with vagal stimulation in attempts to improve mood, performance, or plasticity has been complementary to mental stress research and extensively studied in vagal stimulation studies with implants and recently with noninvasive vagal stimulation tools. Due to differences in stressors, their duration, or the application time of stimulation, the physiological outcomes have been variable, specifically in noninvasive studies. The onset time analysis from this randomized, double-blind trial could be instrumental to the design of clinical studies.

The findings of this study are also salient in the design of next-generation engineering tools to improve quality of life based on n-VNS. Recent advances in noninvasive technologies pave the way towards closed-loop systems that combine wearable sensing and stimulation. Real-time decision making through noninvasively measured parameters stands as the key mechanism for such technologies. This study lays the groundwork for closed-loop n-VNS systems that could provide “dosage” recommendations, personalized optimal delivery, or determination of response time.

**Limitations**

The sample size was relatively low; however, the two groups had comparable baseline characteristics. The onsets of action were manually annotated, as the sample size was low. To ensure reproducibility of results, multiple annotators independently annotated, and the inter-annotator agreement was quantified. The current study is also limited by left vagus stimulation per the study design. The findings should be re-evaluated for trials that include right vagus stimulation, as it is predicted to have different effects on cardiovascular and peripheral function.

**Conclusion and Future Work**

Wearable cardiovascular and peripheral biomarkers—HR, PEP, and PPG amplitude—were modulated in 18 ± 7 seconds for n-VNS without stress and in 16 ± 9 seconds for n-VNS following traumatic stress with a significant difference in the absent onset counts between active and sham groups. The utility of this work is to eliminate ambiguity of the physiological outcomes for n-VNS studies. The findings of this study have applications in clinical studies that use n-VNS in tandem with or without stress, as well as in the design of wearable systems that combine sensing and stimulation. For instance, stimulation timing appears to be important in the outcomes of multiple studies that test the differences in cognitive functioning, memory functioning, psychomotor functioning, or executive functioning with VNS application. The outcomes of this study could be used to design effective clinical studies for n-VNS devices. From the wearable sensing standpoint, the results could be instrumental for decision making algorithms, such as the determination of stimulation timing or the effectiveness of the stimulation. The inclusion of the sham group and multi-day protocol presented in this study provides unique, continuous wearable sensing data to consider for physiological outcomes of similar studies.

Future work should assess approaches to automate this process as more patients are recruited who have been diagnosed with psychiatric conditions. Further investigation into translating these findings from laboratory settings to wearable devices will help generalize this work, and confirming the reproducibility of these results in larger sample sizes is also pertinent. Of particular importance is the sexually dimorphic nature of cardiac and neurological responses in men and women; as the recruitment process continues, we will leverage larger sample sizes in future work to separate based
on sex and quantify any gender differences. The multiparametric determination presented herein may answer questions related to identifying physiological changes during n-VNS and the optimal time of delivery in response to a detected event.

Acknowledgments
This work is based on material supported by the Defense Advanced Research Projects Agency (DARPA), Arlington, VA, under Cooperative Agreement N66001-16-2-4054.

References
Enhancing Multi-Center Patient Cohort Studies in the Managing Epilepsy Well (MEW) Network: Integrated Data Integration and Statistical Analysis

Xinting Hong, BS1, Chang Liu, BS1, Hasina Momotaz, PhD2, Kristin Cassidy, MA2, Martha Sajatovic, MD2, Satya S. Sahoo, PhD1

1Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH
2Department of Neurology, University Hospitals Cleveland Medical Center, Cleveland, OH

Abstract

Self-management techniques that assist patients with chronic conditions, such as epilepsy, diabetes, and arthritis, play an important role in managing and caring for their conditions. The US Center for Disease Control and Prevention (CDC)-funded Managing Epilepsy Well (MEW) Network consists of 11 study sites across the US that aims to develop and disseminate self-management techniques for epilepsy patients. Epilepsy affects more than 65 million patients worldwide with serious negative impact on their own as well as their family member’s quality of life. Taking advantage of advances in biomedical informatics, the MEW Network has created an integrated database (MEW DB) using a common data model and two tiers of study variables. The MEW DB consists of 1680 patient data records covering a wide range of patient population nationwide. Therefore, there is growing interest in the use of the MEW DB for different cohort query analysis. To address the challenges in: (1) selecting appropriate MEW research studies based on inclusion/exclusion criteria; (2) creating a patient cohort for given research hypothesis; and (3) performing appropriate statistical tests; we have developed an integrated data query and statistical analysis informatics tool called Insight. The Insight platform features an intuitive user interface to support the three phases of study selection, patient cohort creation, and statistical testing with the use of an epilepsy domain ontology to support ontology-driven query expansion. We evaluate the Insight platform using four user evaluation methods of “first click testing” and “user satisfaction survey”. In addition, we performed a time performance test of the Insight platform using four patient datasets and three statistical test. The results of the user evaluation show that Insight platform is strongly approved by the users and the results of the time performance show that there is marginal difference in performance as the volume of patient data increases in the MEW DB.

Introduction

Epilepsy is a serious neurological disorder and it severely affects the quality of life of both the patients as well their family members (1, 2). Epilepsy affects more than 65 million persons worldwide with more than 2.8 million persons in the United States who experience seizures, which lead to disabilities, decreased ability for daily functioning, and common comorbidities including depression as well as cognitive impairment (1, 3). The cost of care for epilepsy is high with total cost burden estimated to be more than $36 billion per year, which disproportionately affects the elderly and minority populations who are economically vulnerable (1). The challenges in managing patients with epilepsy are significant with a growing population of older persons (estimated to increase to approximately 80 million by 2060) (4), high utilization of healthcare services often due to non-adherence to medication, and stress on patients as well their family members.

Self-management techniques have been successfully developed for many chronic conditions, including heart diseases, asthma, and depression, which enable patients to address disease symptoms, stress, and improve their quality of life (5). Epilepsy self-management techniques aim to address the limitations of existing approaches for providing care, for example difficulty in monitoring of daily medication adherence, and lack of availability of time or clinical resources to common comorbidities of depression or other psychosocial conditions (6). Self-management techniques for epilepsy patients involves improving their medication adherence, memory, and quality of life (7). Therefore, in 2007, the US Center for Disease Control and Prevention (CDC) created a network of multiple institutions to develop “self-management” techniques for epilepsy called Prevention Research Centers’ Managing Epilepsy Well (MEW) and disseminate the best practices to patient population (8).

The MEW Network institutions have developed multiple epilepsy self-management programs that use multi-modal communication and dissemination techniques to advance the adoption of self-management techniques in the patient population (7). These programs include the Epilepsy Awareness Support and Education (WebEase) project (9), the...
As part of the MEW Network initiative, we have harmonized the data from different MEW Network sites using a traditional Extract Transform Load (ETL) approach to create the MEW database (MEW DB) consisting of 1680 patients from 11 study sites. The MEW DB is being used to analyze different aspects of epilepsy treatment, including self-management techniques. For example, a study by Friedman et al. examined depressive symptoms and suicidality in epilepsy patients who were enrolled in self-management studies (12). The study found that depression and suicidal thoughts are common in epilepsy patients who were enrolled in self-management studies. Similarly, a study by Begley et al. evaluated the demographic and clinical correlated in epilepsy patients to characterize their self-management skills (13). This study used data from 436 patients involving five MEW Network studies and found that there was significant variation in the self-management competencies of patients. These studies highlight the need for well-defined data integration techniques as well as statistical analysis methods that can be used to both create a national-level patient population resource as well analyze it to derive insightful results for potential interventions.

There has been significant amount of work in the informatics community related to data integration for patient cohort query using ontology-driven techniques (14-16). However, to the best of our knowledge, many of these existing data integration platforms have limited or no support for the statistical analysis methods that need to be performed on a patient cohort data after it has been extracted from the integrated database. Users often have to download and perform tedious data transformation as well as curation steps over the patient cohort data before it can be analyzed using appropriate statistical methods such as t-test, Chi-Square, Mann-Whitney U test, ANOVA test, Fisher's exact text and Kruskal-Wallis test. In addition, users often have to rely on external, third-party tools that support these statistical methods for analysis. The lack of support for statistical data analysis in informatics tools for patient cohort query becomes a significant rate limiting step as the number of research studies using the data rapidly increase, for example the governing body for the MEW DB receives multiple requests for new research studies using this unique resource. In addition, the lack of integrated statistical data analysis tools in a data integration platform makes it difficult for users to perform exploratory data analysis by modifying study parameters based on results of statistical analysis.

In our previous work, we developed an epilepsy-focused, ontology-driven data integration and patient cohort query platform called Insight (17). The Insight tool uses the Epilepsy and Seizure Ontology (EpSO) to support data processing, integration, and query to allow users to create research study cohorts via an intuitive visual interface. The results of the cohort queries can be exported as comma separated value (CSV) files for subsequent processing and analysis using statistical tools. In this paper, we describe the development and application of a dedicated statistical analysis component in the Insight platform that can support a variety of statistical methods. This integrated statistical analysis feature of Insight significantly reduces the challenges for patient cohort analysis in the MEW network and we believe it can be easily extended to other domains given the widespread use of statistical data analysis tools in the biomedical research domain. To validate the integrated statistical feature of the Insight platform, we replicate three research studies using MEW-DB data.

Materials and Method

Setting

The goal of the MEW Network is to develop self-management intervention techniques that can be adopted by persons with epilepsy in their home and work environments using a community-based approach (7). At present, the MEW Network consists of centers at eight institutions across the US, including the University of Arizona, Morehouse School of Medicine, New York University, the University of Washington, the University of Minnesota, the University of Illinois at Chicago, and Case Western Reserve University. The MEW Network as developed and tested a wide range of intervention techniques for self-management techniques with a special focus on underserved and minority populations (10, 18-21). The MEW Network works closely with national organizations focused on epilepsy, including the American Epilepsy Society and the Epilepsy Foundation to disseminate evidence-based self-management interventions. This paper integrates and analyzes data in the MEW-DB from 11 studies with a total of 1680 patients and this study was approved by the University Hospital Cleveland Medical Center Institutional Review Board (IRB).

Data Collection
The data in the MEW-DB were obtained as part of the 11 research studies and included a variety of details that based on a set of common data elements (CDE) that are approved by the MEW DB steering committee for inclusion in the integrated dataset. The current set of MEW CDEs are categorized into Tier-1 and Tier-2 variables with Tier-1 variables consisting of 16 CDEs and Tier-2 variables consisting of 15 CDEs. These variables include description of patient demography and clinical outcomes (more details about the variables is presented in our earlier work (17). Table 1 describes the details of the patient data in the MEW DB with a higher percent of female patients (45%) as compared to male patients (28%). The patient data table shows that the MEW DB studies focus on low income group (18% earn less than $25,000 per year).

Reference Studies

The three reference studies used in this paper were conducted using the MEW-DB data. We briefly describe these three studies. The first study (Study 1) used the integrated data to evaluate the mood symptoms in 770 patients who are enrolled in self-management interventions using the 9-item Patient Health Questionnaire (PHQ-9) scores (12). The study found that 43.4% of the patients in the study have moderate to severe depression and about 20.1% of the patients had suicidal ideation. The study used multiple statistical methods to evaluate the correlation between depression variables and PHQ-9 total scores, for example Spearman’s correlation was used to characterize the relation between age and PHQ-9 scores. Similarly, multiple logistic regression was used to characterize the independent association between demographic and clinical variables to PHQ-9 total scores. This study showed that depression and suicidal ideation are common in the patients enrolled in the various self-management studies.

The second study (Study 2) analyzed data from 459 patients that were collected across four research centers to identify correlates of quality of life (QOL) (7). In particular, the study used various demography data, such as age, gender, race, education and employment, together with quality of life QOLIE-10 scale and PHQ-9 scale for depression. The study found that association between the severity of depression and lower QOL scores. The study used Kruskal-Wallis nonparametric methods for analysis of the data. Finally, the third study (Study 3) assessed the outcome of 5 randomized control trials (RCTs) with respect to symptoms of depression in epilepsy patients (13). The study used data from 453 patients with 232 in self-management intervention and 221 patients in usual or wait-list control. Using a series of t-test, the study characterized the changes in PHQ-9 scores over a period of time and a mixed model analysis was done to evaluate the association between baseline variables and changes in PHQ-9 scores over time.

Insight Epilepsy Informatics Platform

The Insight architecture consists of a set of modules that support three primary functions: (1) study selection based on inclusion/exclusion criteria, (2) cohort query formulation using a domain ontology-driven visual query interface, and (3) integrated statistical analysis. The current version of Insight has been developed using the Django web application framework (the previous version of Insight used the Ruby on Rails (17)). The use of Django web framework uses the

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age - Mean(SD)</td>
<td>39.96 (13.99)</td>
</tr>
<tr>
<td>Gender-N (%)</td>
<td></td>
</tr>
<tr>
<td>Female-N (%)</td>
<td>763 (45.42%)</td>
</tr>
<tr>
<td>Male-N (%)</td>
<td>486 (28.93%)</td>
</tr>
<tr>
<td>Race-N (%)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>832 (49.52%)</td>
</tr>
<tr>
<td>Black/African American</td>
<td>176 (10.48%)</td>
</tr>
<tr>
<td>Other</td>
<td>100 (5.95%)</td>
</tr>
<tr>
<td>Ethnicity — N (%)</td>
<td></td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>711 (42.32%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>107 (6.37%)</td>
</tr>
<tr>
<td>Education — N (%)</td>
<td></td>
</tr>
<tr>
<td>High school or less</td>
<td>323 (19.23%)</td>
</tr>
<tr>
<td>At least some college</td>
<td>760 (45.24%)</td>
</tr>
<tr>
<td>Income — N (%)</td>
<td></td>
</tr>
<tr>
<td>&lt;$25K</td>
<td>305 (18.15%)</td>
</tr>
<tr>
<td>$25–50K</td>
<td>83 (4.94%)</td>
</tr>
<tr>
<td>&gt;$50K</td>
<td>166 (9.88%)</td>
</tr>
<tr>
<td>Marital status — N (%)</td>
<td></td>
</tr>
<tr>
<td>Married or partnered</td>
<td>328 (19.52%)</td>
</tr>
<tr>
<td>Other</td>
<td>507 (30.18%)</td>
</tr>
<tr>
<td>30-day seizure frequency — Mean (SD)</td>
<td>5.26 (20.34)</td>
</tr>
<tr>
<td>QOLIE-10 — Mean (SD)</td>
<td>2.89 (0.81)</td>
</tr>
<tr>
<td>PHQ-9 — Mean (SD)</td>
<td>9.35 (6.62)</td>
</tr>
</tbody>
</table>
Python programming language, which has a large number of libraries and modules that support variety of data processing and analysis tasks, including statistical methods. The architecture of Insight allows it to integrate and manage data from multiple studies with the use of either software modules for data pre-processing and conversion to a common data model following the well-known Extract Transform Load (ETL) approach. We note that as part of the data processing step, multiple terms in the study data are annotated using an epilepsy domain ontology, which enables Insight to support ontology-driven query execution.

Figure 1 shows the current architecture of Insight that uses the Model View Template (MVT) approach with data accessed via an object relational data Model, the Insight user interface supported by the View component, and the user interaction with various features of the software mediated by the Template [Django]. The Insight platform is accessed via a Web browser with role based access control (RBAC) with users assigned to different user groups. The access privileges of a user are based on the specific study protocol that is approved by the MEW DB steering committee, which reviews and accepts or declines a proposed study to access the integrated datasets. Users follow a three-step process to identify specific research studies based on their inclusion-exclusion criteria, followed by creation of a study cohort using an intuitive query formulation process, and the study cohort data in finally analyzed using appropriate statistical methods in the data analysis module.

During the first step for study selection, users can select appropriate inclusion and exclusion criteria from a drop-down menu and the specific research studies that conform to the selected inclusion/exclusion criteria are listed for users to review. Users can also choose to select all studies in the MEW-DB for the second phase of cohort query. During the second phase, users can compose a patient cohort query using any of the Tier-1 and Tier-2 variables (described in the previous section above) using a drop-down menu and assign specific values to query variables, for example age between 20-40 years or Patient Health Questionnaire-9 (PHQ-9) score of -5. The selected variables, which are mapped to the Epilepsy and Seizure Ontology (EpSO) (22), are used to compose the patient cohort query followed by “query unfolding”. The query unfolding uses ontology reasoning to include all the subclasses of a term in the user-defined query, which ensures that the query results are exhaustive. EpSO is a domain ontology that has been developed to model multiple aspects of epilepsy, including seizure semiology, medication, and etiology (for detailed description of EpSO we refer to (22)). Users can view the results of the cohort query in the Insight visual interface or download it for statistical data analysis.

However, there are many challenges associated with use of statistical data analysis after downloading the patient cohort data, including re-formatting the data into a format that is required by a given statistical tool (e.g., Mann-Whitney U test and chi-squared test). In addition, downloading and transferring the patient cohort data between different computing platforms that have relevant statistical data analysis tools is cumbersome and significantly increases the time required for statistical analysis. An intuitive approach to address this challenge is to enable users to perform various statistical analysis in the Insight platform itself, which would eliminate the need to download and process the patient cohort data. Therefore, we implemented a new integrated statistical data analytics module in Insight to streamline and effectively reduce the time required to analyze patient cohort data.
Figure 2: The three components of the Insight platform, namely: (1) Study Selection; (2) Patient Cohort Selection; and (3) Statistical Analysis component.
Statistical Data Analytics Module Development

The Insight platform supports multiple statistical methods that can be applied to the patient cohort data. To apply statistical methods on the patient cohort data, a new statistical data analysis module was developed and implemented. Using the patient cohort data after execution of the two phases of Study Selection and Cohort Query, which have been described in details earlier (17), the statistical analysis module uses a similar user-friendly approach for invoking the a set of statistical methods on the results of the cohort query.

The implementation of the statistical methods involves three aspects: (1) storage and modification of data in a common model; (2) user interface features for invocation of statistical methods; (3) visualization of results of statistical data analysis. In the first phase, the results of the patient cohort query are stored as a “view” in the Insight database, which allows fast access to the data during subsequent statistical data analysis. In addition, the database view is also used to support download functions for users. In the second phase, the Insight statistical data analysis module interface is used to conduct the statistical analysis. The statistical functions in Insight are implemented using a modular approach, which allows new statistical modules to easily added. At present, Insight supports multiple statistical analysis methods, including variance, t-test, chi-squared test, and Mann-Whitney U test. As discussed earlier, these statistical methods are used in all the three research studies involving use of the MEW-DB integrated data.

Figure 2 shows the creation of the patient cohort using the first two components of the Insight platform (Study Selection and Cohort Query). Figure 2 shows the application of statistical methods on the patient cohort data using the Insight interface features, where a user can select specific statistical method to be applied to the data. For a given statistical method, the Insight interface automatically populates the subcategories of the drop-down menu with appropriate variables selected from the patient cohort data (output of the second component of Insight described earlier). This adaptive population of drop-down menu allows users to quickly select cohort-specific variables to be analyzed using a statistical method, which is both efficient as well as user-friendly. Once the statistical method and appropriate study variables are selected and applied, the output results are listed for viewing by the user. The results of the analysis can also be downloaded also by the user.

Evaluation of Data Analytics in Insight

The objectives of our evaluation are two-fold: (1) the first goal is to evaluate the usability of the Insight platform using two approaches, namely a “user satisfaction survey” and “first click testing”; and (2) the second goal is to evaluate the time performance of the statistical modules with respect to different datasets. The usability evaluation of the Insight platform was conducted with two team members of the MEW project who have performed several tasks on the MEW DB, including data processing and statistical analysis. We propose to perform a comprehensive user evaluation study of the Insight platform with a focus group consisting of participating MEW network members in the future. The details of two methods used for usability evaluation are:

1. **First click testing**: This method is used to evaluate the effectiveness and efficiency of the user interface in terms of enabling users to accomplish their tasks. We defined four tasks for users to complete by using the Insight user interface for the statistical data analysis, namely: (a) computation of t-test over a select cohort; (b) computation of Mann-Whitney U test over specific study variable; (c) computation of mean; and (d) computation of variance of study variables in a given cohort. This testing measures the number of clicks performed by the users in completing a given task, which may involve selecting the correct or incorrect set of options in the user interface. Intuitively, users who select the correct sequence of options in the user interface complete their tasks with a minimal number clicks and on time.

2. **User satisfaction survey**: In addition to the first click test, we also performed a broader user satisfaction survey with the objective of evaluating: (a) whether users can locate relevant information easily; (b) whether the results of their queries are easily accessible as well as visualized; (c) a comparative evaluation of the Insight interface with respect to existing tool; and (d) the likelihood of users recommending the Insight platform to other users for cohort data analysis.

The response of the users was recorded on a graded scale of 1-10 points with 1 corresponding to lowest value and 10 corresponding to highest value. The performance evaluation test used data from all the three studies that used the MEW DB data (described earlier), which demonstrate the scalability of the Insight platform with different size of datasets.

**Results**

The results are shown in Figure 3 show the results of the two user evaluation surveys for the Insight platform.
First click testing. The results for the first click testing (Figure 3(A)) show that both the users performed extremely well in terms of selecting the correct options in the interface to complete the four tasks (described in the previous section above). The first question asked the users whether they were able to complete the four tasks in the first attempt and both the evaluation results show that both the users completed the tasks in their first attempt. The second question validated the design principle of the user interface that users should be able to reach a final selection option with minimal number of clicks. Both the users agreed strongly (user rating score of 10) in the third question that the drop-down menu showed only the relevant study variables for a give statistical data analysis method. The two users gave a rating of 10 and 9 (corresponding to user 1 and user 2) in the fourth question for the difficulty level associated with locating a specific statistical data analysis method in the user interface. Finally, both the users agreed that the Insight platform was very easy to use to perform statistical analysis as compared to existing statistical packages such as SAS. However, one of the users noted that at present the Insight platform did not support a large number of statistical data analysis methods, which we aim to address in the future with implementation of additional statistical data analysis methods.

<table>
<thead>
<tr>
<th>Insight: First Click User Evaluation</th>
<th>Insight: User Satisfaction Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="User Evaluation Graph" /></td>
<td><img src="image2" alt="User Satisfaction Survey Graph" /></td>
</tr>
</tbody>
</table>

**Figure 3:** The results of the user evaluation of the Insight platform are shown using two approaches: (A) First Click user evaluation; and (B) User satisfaction survey with two users who are familiar with the MEW DB

User Satisfaction Survey. Figure 3(B) shows the results of the user satisfaction survey consisting of 5 questions covering the navigation, visualization, effort required to complete tasks, and whether they would recommend the Insight platform to other MEW DB users. One of the users noted that the Insight platform was “fairly easy (to use) after demonstration” and it would have been more difficult to use without a demonstration. We agree with this assessment and we aim to create comprehensive user manual for the Insight platform that will allow first time users to quickly learn the different components of the platform. Both the users agreed that it was relatively easy to select specific statistical analysis method the associated study variable in the second question. Similarly, both the users strongly agreed that results of the three Insight platform

| ![Performance Graph](image3) |

**Figure 4:** A comparative evaluation of the time performance of the Insight platform using four different sizes of patient data and three statistical test.
components (Study Selection, Cohort Selection, and Statistical Analysis) were clearly visualized in the third question. Although both the users strongly agreed that the Insight platform reduces the time and effort required to conduct statistical data analysis as compared to existing methods in the fourth question, one of the users noted that at present the Insight platform does not allow for more in-depth data analysis. We aim to work with the users to address this concern. Both the users strongly agreed in the fifth question that will recommend the Insight platform to other MEW DB users.

Performance evaluation for statistical analysis. To demonstrate the practical usability and scalability of the Insight platform, we used four datasets consisting 100, 200, 400, and 450 patients from Study 3 to compute three statistical tests, namely t-test, Chi-Squared test, and Mann-Whitney U test. Figure 4 shows the time performance of the Insight platform statistical analysis module in milliseconds for the four datasets and three statistical tests. The results show that performance difference between the four datasets is marginal for all the three statistical tests. For example, there is a difference of 0.2 milliseconds and 0.1 milliseconds between the 100 patient datasets and the 200 as well as 400 patient datasets respectively. Similarly, there is a difference of 0.1 to 0.4 milliseconds for the t-test. These results show that there is no significant impact on the time performance of the Insight platform in the MEW-DB as the number of patients increases. It is also interesting to note that the difference between the time performance of the three statistical tests is also marginal with difference in values ranging from 0.1 milliseconds (for 100 patients) to 0.7 milliseconds (for 200 as well as 450 patient datasets).

Discussion and Conclusion

This study demonstrated the significant utility of a focused ontology-driven informatics tool for epilepsy clinical research that offers functionalities spanning study selection, cohort identification, and integrated statistical analysis in the context of a multi-center national study on epilepsy self-management. We are not aware of other existing tools that offer a similar level of granularity and functionality that can be readily used in the MEW Network project. The utility of the Insight platform and in particular the statistical data analysis module lies in its potential to significantly improve the efficiency and effectiveness of data analysis procedure for users as the size of the MEW DB continues to increase in terms of the volume of data and the number of studies proposed to be conducted on the integrated data. Although there are multiple data integration and ontology-driven informatics tools that are available in the biomedical research domain, there are significant challenges associated with their adoption in highly specialized, specific disciplines of medicine. Therefore, we decided to develop the Insight platform as an epilepsy-focused informatics tool that meets the requirements of this domain with the use of a domain-specific ontology (EpSO) and specialized data processing, querying, and analysis features.

The availability of integrated functionalities, in particular the statistical analysis module in the Insight platform has several advantages as compared to the use of offline statistical analysis tools. For example, users do not need to download and transfer data across one or more computers that increases the risk to safe management of research study data, including de-identified study data as used in the MEW-DB. Therefore, the integrated functionalities for statistical testing in the Insight platform offers significant advantages in terms of ease of use as well as maintaining data access restrictions for users as compared to existing approaches of manual data transfer and analysis.

The primary objective of the Insight platform is to serve as an informatics tool that facilitates greater use of the MEW DB for analysis and evaluation of self-management techniques in epilepsy. To achieve this goal, we are implementing new features in the Insight platform, including the development of a Docker-based version, which will allow the Insight platform to be easily shared and installed by the MEW Network community. The Docker-based version will allow individual MEW researchers

The work reported in this paper has some limitation. For example, as highlighted by a user during the user evaluation the current version of the Insight platform does not support in-depth statistical analysis. In addition, the number of statistical tests currently available in the Insight platform are limited. Therefore, we plan to significantly increase the number of statistical tests available in the next version of the Insight platform. In addition, we propose to add extensive result visualization features in the next version of the Insight platform.

In conclusion, in this paper we presented an integrated data processing and analysis platform called Insight that can significantly facilitate epilepsy patient cohort studies in the MEW Network in the context of epilepsy self-management studies.

Acknowledgement. This work is supported in part by the CDC grant 1U48DP005030, NIH Big Data to Knowledge (BD2K) grant (1U01EB020955), and NSF grant# 1636850.
References

5. CDC U. Self Management Education for Chronic Conditions.
Trends and characteristics of protected health information breaches in the United States

Md Mahbub Hossain, MBBS MPH,1 Y. Alicia Hong, PhD2

1Department of Health Promotion and Community Health Sciences, School of Public Health, Texas A&M University, TX; Department of Health Administration and Policy, College of Health and Human Services, George Mason University, VA.

Abstract

Objectives: To evaluate the data breaches incidents in the U.S. between 2010 and 2018, identify the characteristics of breaches involving more than a million records, and compare the changes before and after wide adoption of EHR in 2015. Materials and methods: Incidents of data breaches between 2010 to 2018 were retrieved from the Office of Civil Rights portal. Descriptive statistical analyses were performed to assess the trends and characteristics, and changes between states from 2015 to 2018 were assessed and mapped. Results: From 2010 to 2018, a total of 2,529 breaches affected 194.74 million individual records. Overall, 72.08% incidents involved healthcare providers; theft (32.94%) and hacking (22.7%) were major types of breaches. Large cases affecting more than a million records happened due to compromised internal structures and systems. After 2015, the magnitude of the data breaches has changed at varying levels in the U.S. states necessitating further research and actions.

Introduction

Use of electronic health records (EHR) offers several benefits including enhanced communication and decision-making in the points of care,2 improve adherence to advised preventive or therapeutic measures by providing timely reminders to the patients or caregivers,4 reduce unnecessary laboratory tests by coordinating previous and prospective diagnostic plans,5 improve quality of care,6 and manage payments and reimbursements.7 The adoption of EHR across different health systems followed diverse pathways.8 Critical challenges like technological complexities, legal issues, economic costs, concerns for safety and security of the records influenced the development of EHR.9 In the United States, two major acts were introduced to address the legal and security-related challenges regarding protected health information.10 Firstly, the Health Insurance Portability and Accountability Act (HIPAA) was passed in 1996 which emphasized physical, administrative, and technical safeguards.10 Secondly, the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 enforced the implementation and utilization of EHR.11 This act mandated the Centers for Medicare and Medicaid Services (CMS) recipients to adopt EHR by the beginning of 2015 to continue receiving full reimbursements and preventing penalties due to inability to adopt EHR.12 In addition, the HITECH Act also stresses the importance of reporting incidents of data breaches involving protected health information.11 Since 2009, the number of providers adopting EHR increased alongside the number of data breach incidents.13 Such incidents erode the trust of individuals on digital health technologies and jeopardize the potential benefits of technological advancements.7,14 In addition, the direct economic burden of data breaches can be as high as $6.2 billion with other legal and administrative consequences.15,16

Several studies have reported the characteristics of those breaches,13,17 types and media locations of breach events,18 geographic locations and types of covered entities,19 and causes behind those incidents in different time periods.20 However, there is a lack of evidence that illustrates the changes in the trends and characteristics of the data breach for all the complete years since the HITECH Act was passed. Moreover, several incidents involved more than a million records per breach,17 but the detailed causes and characteristics of those large cases are not explored yet. Furthermore, the number of breach incidents or affected records might not provide the actual picture of data breach in the U.S. at the state level due to varying population size from state to state. It is critical to examine the population adjusted individual records affected due to data breaches after the mandatory adoption of EHR across the nation since 2015. The objectives of this study are 1) identify all the reported incidents of data breaches in the U.S. between 2010 and 2018, 2) list the key characteristics of incidents which involved more than a million individual records, and 3) describe
the changes in population adjusted records affected in different U.S. states following the mandatory adoption of EHR in 2015.

Materials and methods

In this study, we used the data of all the reported breach incidents between 2010 to 2018 from the breach portal of the Office of Civil Rights at the U.S. Department of Health and Human Services. According to section 13402(e)(4) of the HITECH Act, the incidents of the data breaches involving protected health information of more than 500 individuals are reported to OCR. We retrieved these publicly available data which provide the name and type of covered entity, the presence of business associates, location, time, and the number of affected individual records for the incidents. Moreover, specific types of the breaches including hacking, improper disposal, theft, loss, and unknown types are reported. We grouped the entries with more than one type of breach as multiple types of breaches in this analysis. Further, the data informs the media location of the data breaches including a network server, paper, desktop, laptop computer, portable devices, email, electronic medical record, and others. In this category, we grouped the entries with more than one media location as multiple locations. Using the above-mentioned data in Microsoft Excel and Stata 15.0 (College Station, TX), we performed descriptive statistics and two-way measures of association using chi-squared tests.

Furthermore, to examine the changes in data breaches following the adoption of EHR across the nation, we calculated the total number of affected individual records at the state level for 2015 to 2018. We adjusted this annual breach data for each thousand population using the annual population estimates for each state derived from the U.S. Census Bureau. We used the population adjusted affected individual records per state to create maps for each year illustrating the changes in the data breach incidents from 2015 to 2018 using ArcGIS software. Moreover, the breach data includes web description of the cases which include qualitative information about the characteristics and consequences of the incidents. We used the description corresponding to the cases that involved more than a million individual records in assessing the characteristics of those large data breach incidents.

Results

Characteristics of data breaches from 2010 to 2018

Total 194.74 million individual records were affected due to 2,529 incidents of the data breaches over a period of nine years from 2010 to 2018 as shown in Table 1. The number of breaches increased every year compared to the previous years except in 2015.

Table 1: A brief overview of the protected health data breaches in the United States (2010-18)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of individual records breached (in million)</td>
<td>194.74</td>
<td>5.93</td>
<td>13.16</td>
<td>2.85</td>
<td>7.02</td>
<td>17.45</td>
<td>113.28</td>
<td>16.66</td>
<td>5.14</td>
<td>13.24</td>
</tr>
<tr>
<td>Number of total breaches</td>
<td>2,529</td>
<td>199</td>
<td>200</td>
<td>218</td>
<td>278</td>
<td>314</td>
<td>268</td>
<td>327</td>
<td>359</td>
<td>366</td>
</tr>
</tbody>
</table>

Types of the covered entity* (percentage)

<table>
<thead>
<tr>
<th>Types of the covered entity</th>
<th>Business associate</th>
<th>Health plan</th>
<th>Healthcare clearing house</th>
<th>Healthcare provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business associate</td>
<td>14.23</td>
<td>13.40</td>
<td>0.16</td>
<td>72.08</td>
</tr>
<tr>
<td>Health plan</td>
<td>22.11</td>
<td>10.55</td>
<td>0</td>
<td>67.34</td>
</tr>
<tr>
<td>Healthcare clearing house</td>
<td>22</td>
<td>9.5</td>
<td>0.5</td>
<td>68.0</td>
</tr>
<tr>
<td>Healthcare provider</td>
<td>2012</td>
<td>6.83</td>
<td>0.46</td>
<td>70.64</td>
</tr>
<tr>
<td>Hacking/IT Incident</td>
<td>22.7</td>
<td>13.06</td>
<td>0</td>
<td>69.42</td>
</tr>
</tbody>
</table>

Type of breach* (percentage)

| Types of breach*                     | 4.02              | 6.83        | 0.46                     | 69.42               | 70.64               | 61.78               | 72.76               | 78.29               | 80.22               | 74.59               |
Improper disposal | 2.93 | 4.02 | 3.02 | 3.21 | 4.32 | 2.55 | 2.24 | 2.14 | 3.06 | 2.46
Loss | 6.17 | 7.04 | 7.5 | 7.8 | 7.19 | 6.37 | 9.33 | 4.89 | 4.46 | 3.55
Multiple types reported | 31.87 | 8.54 | 20.5 | 22.48 | 29.14 | 36.94 | 38.06 | 39.76 | 35.38 | 39.07
Other | 2.97 | 10.55 | 1.5 | 5.96 | 5.76 | 7.01 | 0 | 0 | 0 | 0
Theft | 32.94 | 65.33 | 56.5 | 55.96 | 42.81 | 35.67 | 29.10 | 18.65 | 15.6 | 11.48
Unknown | 0.4 | 0 | 3.5 | 0 | 0.72 | 0.32 | 0 | 0 | 0 | 0

| Media location of breached data* (percentage) | Desktop computer | 6.33 | 10.55 | 10 | 10.55 | 11.51 | 5.1 | 4.48 | 4.28 | 2.51 | 3.55
| Email | 13.33 | 2.01 | 0.5 | 4.13 | 7.55 | 10.51 | 12.69 | 12.23 | 23.96 | 29.78
| Laptop | 12.65 | 24.12 | 17 | 22.94 | 23.02 | 12.1 | 11.57 | 6.73 | 4.46 | 4.64
| Multiple locations | 21.43 | 21.61 | 28 | 19.27 | 17.27 | 21.66 | 24.25 | 22.02 | 23.4 | 17.49
| Network server | 16.13 | 9.05 | 8.5 | 9.17 | 11.87 | 17.83 | 14.93 | 24.46 | 23.4 | 16.39
| Other | 8.54 | 9.55 | 13 | 11.01 | 8.27 | 10.83 | 5.97 | 7.95 | 5.57 | 7.65

* p <0.001

the highest number of individual records were breached (n=113.28 million) in 2015 (Figure 1). Moreover, data breaches varied with the types of covered entities which was statistically significant. Most incidents (72.08%, n=1,832) happened in covered entities which were healthcare providers followed by the business associate (14.23%, n=360) and health plan (13.4%, n=339). Further, different types of breaches were significantly associated with the number of breaches each year. Overall, theft (32.94%) and hacking or IT incident (22.7%) were major types involved in breach incidents whereas a significant proportion (31.87%) of breaches involved more than one type. However, hacking or IT incidents were much lower (4.02%) in 2010 which had increased in 2018 (43.44%).

Figure 1: Number of incidents and individual records affected due to data breaches in the U.S. between 2010 and 2018.
Similarly, the number of multiple types of breaches were reported much lower in 2010 (8.54%) compared to 2018 (39.07%). In addition, the breaches happened through different media; papers or films (21.59%) and multiple media locations (21.43%) were reported alongside the breach incidents. Also, the rate of incidents happened through laptop had declined from 24.12% in 2010 to 4.64% in 2018. In contrast, incidents through network server had increased from 9.05% in 2010 to 16.39% in 2018. Similarly, breach incidents happened through email was only 2.01% in 2010 which increased up to 29.78% of annual incidents in 2018.

**Breaches affecting more than a million records at a time**

Among 194.74 million total breached records, only 23 incidents involved more than a million records each, resulting in a total 151.81 million records which are about 78% of all records breached between 2010 and 2018 (Table 2). More than 105 million breached records were associated with health plans representing 69% of all large incidents of breached data. At the state level (Appendix-1), Florida, New York, and Tennessee had a higher number of such large incidents (n=3 for each of these states).

**Table 2: Summary of breaches affecting more than a million records**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Type</th>
<th>Number of breach events</th>
<th>Number of records affected (in million)</th>
<th>Percentage of all breaches that involved million records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of total breaches</td>
<td>All types</td>
<td>23</td>
<td>151.81</td>
<td>100%</td>
</tr>
<tr>
<td>Covered entity</td>
<td>Business associate</td>
<td>10</td>
<td>30.94</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>Health plan</td>
<td>8</td>
<td>105.45</td>
<td>69%</td>
</tr>
<tr>
<td></td>
<td>Healthcare provider</td>
<td>5</td>
<td>15.42</td>
<td>10%</td>
</tr>
<tr>
<td>Type of breach</td>
<td>Loss</td>
<td>2</td>
<td>5.95</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>Multiple types</td>
<td>2</td>
<td>3.25</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Theft</td>
<td>5</td>
<td>12.47</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>Hacking/IT incident</td>
<td>13</td>
<td>128.24</td>
<td>84%</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>1</td>
<td>1.90</td>
<td>1%</td>
</tr>
<tr>
<td>Media location of breach</td>
<td>Desktop</td>
<td>1</td>
<td>4.03</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>Email</td>
<td>1</td>
<td>1.42</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Laptop</td>
<td>1</td>
<td>1.22</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Multiple location</td>
<td>4</td>
<td>11.22</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>Network server</td>
<td>11</td>
<td>123.79</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>5</td>
<td>10.13</td>
<td>7%</td>
</tr>
<tr>
<td>Status of the cases in January 2019</td>
<td>Under investigation</td>
<td>3</td>
<td>5.32</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>No description reported</td>
<td>11</td>
<td>46.44</td>
<td>31%</td>
</tr>
<tr>
<td></td>
<td>Investigation completed and description reported</td>
<td>9</td>
<td>98.05</td>
<td>65%</td>
</tr>
</tbody>
</table>

Most of the breaches (n=13 out of 23) occurred through hacking or IT incident resulting in 128.24 million breached individual records. In addition, most of the incidents (n=10) occurred using network servers whereas 4 incidents involved multiple media locations. Among all 23 breaches affecting millions, only 9 had web descriptions reported in the OCR data breach portal. Five out of 9 such incidents involved hacking or IT incidents as the type of breaches and network server as the medium of the breach. The largest incident affecting the highest number of individual records (n=78.8 million) occurred in 2015 through a series of cyber-attacks to the IT system of the covered entity. Other cases affecting millions of individuals involved unauthorized access (n=3.47 million), breaching SQL database (n=2.21 million), hacking computer records (n=4.5 million), misuse of information systems through hacking (n=1.1 million), displaced locked cabinet (n=1.05 million), theft of unencrypted tapes (n=1.7 million), stolen laptops (n=1.22 million), and multiple breaches due to unprotected systems and processes (n=4.02 million). Three cases reported settlements ranging from $2.3 to $16 million paid to OCR as penalties whereas all cases reported that the covered entities implemented varying protective measures to prevent data breach incidents.
Changes in health data breaches at the state level from 2015 to 2018

Population-adjusted rate of affected individual records illustrated varying severity of data breaches across the U.S. states (Figure 2). The overall adjusted rates show a decline in data breaches from 2015 to 2018 with different rates of individual records affected in different states in different times.

In 2015, Indiana had the highest rate of individual records breached (1.26 million affected records per 100,000 population) followed by Washington, New York, Maryland, and California. In the subsequent years, Arizona (65,140 affected records per 100,000 population), Kentucky (16,438 affected records per 100,000 population), and Arkansas (7,129 affected records per 100,000 population) had the highest population-adjusted data breaches in 2016, 2017, and 2018 respectively. Moreover, Washington, New York, California, and Georgia had a higher population adjusted affected records in all four years. In contrast, Idaho, Hawaii, North and South Dakota had lower population adjusted affected records throughout the study period.

Figure 2: Changes in the population-adjusted breached records from 2015 to 2018 in the United States except Hawaii, Alaska, and Puerto Rico.
Discussion

In this study, we evaluated all reported incidents of data breaches in the U.S. covering all the completed years after 2009 when the data breach registry was introduced under the HITECH Act. Our analysis shows a high burden of the data breach from 2010 to 2018 with a varying number of incidents and individual records affected. In addition, the changes in the covered entity, types of breaches, and media location of data breaches provide a broader picture of data breaches in the U.S. Moreover, the population-adjusted data breaches in 2015 after mandatory adoption of EHR illustrates the severity of data breaches in most of the states, which had decreased in subsequent years. However, many states continue to suffer from a high rate of the data breach, which is a major public health concern. Furthermore, the large data breaches affecting more than a million records per incident offer several critical insights. First, more than 72% breach incidents involved healthcare providers; in contrast, 8 large incidents involved health plans resulting in more than 105 million individual records, which is attributable to 69% of large breaches and 54% of all breaches. Therefore, careful attention should be given to the types of entities covered under HIPAA for securing the protected health information. Second, the increased rate of data breaches through hacking or IT incidents affecting 84% of the major breaches highlights the necessity of safeguarding the EHR systems. In addition, more than 123 million records were breached through the network servers, which revalidates the high number of IT-related incidents. Third, among the 23 major incidents, 11 resolved cases involving more than 46 million cases did not have any description on the OCR portal. A substantial lack of such information can hinder scientific, administrative, economic, and legal analyses about the respective cases.

To mitigate the burden of data breaches in the U.S., it is critical to examine the reasons for such events and explore intervention strategies to address the same. First, at the institutional level, a lack of effective resources is a major challenge to prevent breaches. Such resources include physical infrastructure that can store and manage health records efficiently, electronic systems protected with adequate authorization processes, and human capital that can effectively utilize the health data management systems without compromising the safety and security of data. For the physical and electronic components, standardizing the inputs and processes may improve the safety outcomes of the health information systems. Second, it is essential to incorporate the data security competencies in the health education and training programs, highlight such competencies in the hiring process, promote the work culture ensuring data safety, and unauthorize the credentials from the information systems while replacing the human resources. Third, patients and their informal caregivers are increasingly using EHR and other digital health platforms. They should be empowered with adequate knowledge, safety practices, and access to secured patient data management systems. This can improve their EHR usage and engage them in preventive measures adopted by the providers and institutions. Fourth, at the systems level, exploring the challenges and opportunities to secure health information is central to bring changes for broader impact. One such critical challenge is the scope of HIPAA which was introduced earlier than the adoption of EHR. In that time, the instruments and technologies were not that complex as they are today. Therefore, revisiting and updating the legal and regulatory measures can strengthen the digital health systems against potential breaches. Fifth, data breaches affecting protected health information can involve federal agencies other than HHS. Recent investigations and settlements conducted by FBI and FTC, assessment of health information products and services by FDA are a few examples of how many agencies can engage during health data breaches. It is essential to define the roles and scopes of these institutions and develop a meaningful partnership to address health data breaches. Lastly, the institutional and systems-level approaches to secure protected health information would require sustained financing. The HITECH Act provides up to $27 billion over 10 years as incentives encouraging the adoption of EHR systems. Leveraging such options and bridging other health systems financing strategies would be vital to the successful prevention of health data breaches in the United States.

This study has several limitations. First, we could not assess the unique individual records out of all the breached records. Without such precision, it cannot be concluded that 194.74 million of breached records belong to the equal number of people. Second, the reported cases contain 500 or more individual records. This implies that all those breaches that had less than 500 records were not reported to OCR and therefore, those incidents are not included in this study. This systematic exclusion of unknown events jeopardizes the actual estimation of data breaches in the U.S. Third, we grouped the types and media locations of data breaches if there were more than one entry. An itemized number of affected records from multiple components could provide a more accurate characterization of the types and media locations of breach incidents. Fourth, we could not find what were the contents of breached protected health
information, which could show the magnitude of how much importance the contents would have carried and how they can potentially be abused against the individuals and other stakeholders. Fifth, it is possible that there were other administrative, financial, or institutional data that were breached alongside the reported cases, which were not primarily covered under HIPAA as protected health information. Such information is not reported to OCR, but implicit data about covered entities serving a specific population in a specific geographic location can be used to derive inferences about health behavior, access, utilization, billing, and a wide range of outcomes. We could not find such data in the OCR portal and examine the same, which is another limitation of this study. Sixth, we could not evaluate if the reported incidents of data breaches, particularly those involving hacking or IT incidents, did make any changes within the data stored in those affected databases. Such manipulations of sensitive data might have more severe and long-term consequences. Lastly, we evaluated only the OCR-reported measures under the HITECH Act. We did not evaluate additional legislative and regulatory measures in the state-level which might have influenced the trends and characteristics of data breaches. Future research should explore these avenues, address the limitations of this study, and inform the development of multi-level interventions to address health data breaches in the U.S.

Conclusion
Advancements in health information technologies have improved the access to health services, however, the raising security concerns of adopting advanced technologies may affect the individuals and providers enormously instead of increasing the efficiency of the existing systems. The characteristics of data breaches and trend over the past nine years inform the high magnitude of this serious public health problem in the United States. Also, repeated breaches in different location and time using diverse strategies and media types indicate the overall weakness of the health system to ensure the safety of protected health information. To address such a massive problem which has been affecting millions of people, rigorous research is essential to explore nature and reasons for the data breaches. The findings of empirical research studies can facilitate the development of evidence-based policies and programs aiming the prevention of future breaches. Last but not the least, strong economic and political commitment is required to implement the multipronged preventive measures at the institutional and systems levels, where the key stakeholders can collaborate protecting the health data safety of the U.S. population.

References


### Appendix 1: Characteristics of data breach incidents that affected more than a million individual records

<table>
<thead>
<tr>
<th>Name of the entity</th>
<th>Place and time</th>
<th>Number of affected individual records (in million)</th>
<th>Description of the incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>AccuDoc Solutions, Inc.</td>
<td>North Carolina, 2018</td>
<td>2.65</td>
<td>Under investigation</td>
</tr>
<tr>
<td>Employees Retirement System of Texas</td>
<td>Texas, 2018</td>
<td>1.25</td>
<td>Under investigation</td>
</tr>
<tr>
<td>Iowa Health System d/b/a UnityPoint Health</td>
<td>Iowa, 2018</td>
<td>1.42</td>
<td>Under investigation</td>
</tr>
<tr>
<td>Newkirk Products, Inc.</td>
<td>New York, 2016</td>
<td>3.47</td>
<td>Unauthorized individuals accessed the electronic protected health information (ePHI) of 3,992,270 members of health plans and claims administrators before acquisition by a new parent company from the older one.</td>
</tr>
<tr>
<td>Banner Health</td>
<td>Arizona, 2016</td>
<td>3.62</td>
<td>Not available</td>
</tr>
<tr>
<td>21st Century Oncology</td>
<td>Florida, 2016</td>
<td>2.21</td>
<td>Two separate incidents breached the SQL database having millions of individual records of the covered entity which agreed to pay $2.3 million in lieu of potential civil money penalties to the Office for Civil Rights (OCR) in U.S. Department of Health and Human Services (HHS) and adopt a corrective action plan to settle probable violations.</td>
</tr>
<tr>
<td>Excellus Health Plan, Inc.</td>
<td>New York, 2015</td>
<td>10</td>
<td>Not available</td>
</tr>
<tr>
<td>Medical Informatics Engineering</td>
<td>Indiana, 2015</td>
<td>3.9</td>
<td>Not available</td>
</tr>
<tr>
<td>University of California, Los Angeles Health</td>
<td>California, 2015</td>
<td>4.5</td>
<td>A hacking incident breached the computer network of the entity which had about 4.5 million individual records. Office for Civil Rights (OCR) collected assurances that corrective measures are implemented following the incident.</td>
</tr>
<tr>
<td>CareFirst BlueCross BlueShield</td>
<td>Maryland, 2015</td>
<td>1.1</td>
<td>Not available</td>
</tr>
<tr>
<td>Premera Blue Cross</td>
<td>Washington, 2015</td>
<td>11</td>
<td>Not available</td>
</tr>
<tr>
<td>Anthem Inc.</td>
<td>Indiana, 2015</td>
<td>78.8</td>
<td>A series of cyber-attach breached the IT systems through phishing emails which affected a database containing about 79 million ePHIs. OCR’s investigations revealed a failure to conduct appropriate risk analysis, review IT systems, respond to security issues and protect ePHIs. A settlement of $16 million was paid to OCR and additional corrective measures were reported.</td>
</tr>
<tr>
<td>Xerox State Healthcare, LLC</td>
<td>Texas, 2014</td>
<td>2</td>
<td>Not available</td>
</tr>
<tr>
<td>Community Health Systems Professional Services Corporations</td>
<td>Tennessee, 2014</td>
<td>4.5</td>
<td>Not available</td>
</tr>
<tr>
<td>Company/Department</td>
<td>Location, Year</td>
<td>Score</td>
<td>Details</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>Community Health Systems Professional Services Corporation</td>
<td>Tennessee, 2014</td>
<td>4.5</td>
<td>Not available</td>
</tr>
<tr>
<td>Montana Department of Public Health &amp; Human Services</td>
<td>Montana, 2014</td>
<td>1.1</td>
<td>A server hacking incident allowed misuse of its information system resources for almost 9 months. This incident affected more than 1 million individuals' personal and health information. After that, several safeguarding and technical enhancement was adopted.</td>
</tr>
<tr>
<td>Advocate Health and Hospitals Corporation, d/b/a Advocate Medical Group</td>
<td>Illinois, 2013</td>
<td>4.02</td>
<td>Multiple breaches were found and OCR reported critical issues about the covered entity including potential risks and vulnerabilities to all of its ePHI, implement adequate policies and procedures to limit unauthorized access to the electronic information systems, involve business associate to safeguard all ePHI in its possession and safeguard unencrypted laptop. The covered entity agreed to pay a settlement amount of $5.55 million and implement a corrective action plan.</td>
</tr>
<tr>
<td>Science Applications International Corporation (SA)</td>
<td>Virginia, 2011</td>
<td>4.9</td>
<td>Not available</td>
</tr>
<tr>
<td>The Nemours Foundation</td>
<td>Florida, 2011</td>
<td>1.05</td>
<td>A locked cabinet was removed from an IT service desk containing the electronic protected health information (ePHI) of 1.05 million individuals. Further, the covered entity improved various safeguards including advanced storage systems, encryption and dual factor authorization. OCR obtained assurances about the corrective measures.</td>
</tr>
<tr>
<td>IBM</td>
<td>New York, 2011</td>
<td>1.9</td>
<td>Not available</td>
</tr>
<tr>
<td>GRM Information Management Services</td>
<td>New Jersey, 2011</td>
<td>1.7</td>
<td>Unencrypted backup tapes of clinical system containing electronic protected health information (ePHI) of 1.7 million individuals were stolen. The involved business associate was terminated from contract and new contract as well as other corrective measures were adopted.</td>
</tr>
<tr>
<td>BlueCross BlueShield of Tennessee, Inc.</td>
<td>Tennessee, 2010</td>
<td>1.02</td>
<td>Not available</td>
</tr>
<tr>
<td>AvMed, Inc.</td>
<td>Florida, 2010</td>
<td>1.22</td>
<td>Two unsecured laptop computers containing ePHI (with demographic and clinical information, diagnoses, lab results, treatment and other data) were stolen from the covered entity's premises. The entity adopted new policies and procedures to safeguard the records and prevent such events.</td>
</tr>
</tbody>
</table>
Health Cards to Assist Decision Making in Consumer Health Search

Jimmy, MIS.\textsuperscript{1,2}, Guido Zuccon, PhD.\textsuperscript{1}, Gianluca Demartini, PhD.\textsuperscript{1}, Bevan Koopman, PhD.\textsuperscript{3}  
\textsuperscript{1}University of Queensland, Brisbane, Australia; \textsuperscript{2}University of Surabaya, Surabaya, Indonesia; \textsuperscript{3}Australian E-Health Research Center, CSIRO, Brisbane, Australia.

Abstract

We investigate the effectiveness of health cards to assist decision making in Consumer Health Search (CHS). A health card is a concise presentation of a health concept shown alongside search results to specific queries. We specifically focus on the decision making tasks of determining the health condition presented by a person and determining which action should be taken next with respect to the health condition. We explore two avenues for presenting health cards: a traditional single health card interface, and a novel multiple health cards interface. To validate the utility of health cards and their presentation interfaces, we conduct a laboratory user study where users are asked to solve the two decision making tasks for eight simulated scenarios. Our study makes the following contributions: (1) it proposes the novel multiple health card interface, which allows users to perform differential diagnoses, (2) it quantifies the impact of using health cards for assisting decision making in CHS, and (3) it determines the health card appraisal accuracy in the context of multiple health cards.

Introduction

It is common practice for people to search the Web for health advice and information about conditions, treatments, experiences and health services – we refer to these search activities as Consumer Health Search (CHS). CHS is a challenging domain where effective search is hindered by vocabulary mismatch and the users’ lack of domain expertise. These issues affect both query formulation and result appraisal\textsuperscript{1}. A study by Zeng et al\textsuperscript{2} showed that, while the general public believes that they were effective in searching for medical advice online, 70% of the study’s participants were relying on incorrect advice. Furthermore, Fox & Duggan\textsuperscript{3} found that 38% of CHS users did not seek professional attention once they found medical advice online. This is problematic as in some cases incorrect medical diagnosis and treatment could lead to a fatal outcomes.

This study investigates the effectiveness of health cards to assist decision making in CHS (e.g., Figure\textsuperscript{1}). Health cards are a specific type of entity card that have recently been introduced by major web search engines to provide quicker access to trusted information relating to a specific health concept\textsuperscript{4}. In general Web search, entity cards are effective in supporting user search activities by presenting heterogeneous information in a coherent way\textsuperscript{5}. Yet, no previous work has thoroughly investigated the effectiveness of health cards to assist decision making in CHS.

We specifically focus on the health decision making tasks of (A) determining the health condition presented by a person (self-diagnosis) and (B) determining which action should be taken next with respect to the health condition (e.g., consult a doctor, self-treat, etc.). To support users in making these decisions, we propose a novel interface that shows multiple health cards within a search engine result page (Figure\textsuperscript{1} right) – we call this multi-cards – in place of traditional single health card interfaces (Figure\textsuperscript{1} left). The multi-cards are inspired by interfaces for product comparison used within shopping websites\textsuperscript{6} (e.g., to compare laptops, shoes, etc.) where features of the compared products are summarised and presented side-by-side to assist the user in their purchase decisions. We believe that the multi-cards would allow CHS users to perform differential diagnoses by quickly, and with less effort, comparing their health observations with several probable conditions at once. In this context, we aim to address the following questions:

RQ1: How do single and multiple cards influence CHS users when making health decisions? The impact of health cards on CHS decisions is measured by: (1) the use of health cards as a source of information; (2) the correctness of decisions made on their basis; (3) the time needed to make decisions; (4) the number of web pages opened; (5) the rate of good abandonment\textsuperscript{7}; and (6) the level of confidence in the decisions.

The multi-cards solution shows a set of health cards rather than a single card. This is useful in contexts where the search systems is unsure about which single relevant card should be shown to the user. The ability of showing multiple health cards increases the chance that the relevant (correct) health card is shown for the user’s condition. However, it
is unclear whether users would be capable of identifying the correct card for their condition. We investigate this in our second question:

**RQ2: How accurate are CHS users in appraising the correctness of health cards?** Search results appraisal is challenging in CHS, and is affected by medical terminology, lack of prior knowledge, and cognitive biases, among others. We investigate if this holds for the use of health cards by measuring how well CHS users identify the correct health cards for their health situation, in the context of a multi-cards interface.

To answer these questions, we conduct a study where 64 participants are presented with eight health scenarios and pre-formulated queries, and are asked to consider a search engine result page (SERP) containing health cards and search result snippets. Participants are then left to interact with the SERP and are asked to make two decisions: (A) What is the most likely health condition for the scenario? (B) What would you do next? Participants are rotated across two interfaces: one with a single health card, and one with the multi-cards.

**Methods**

A within-subjects user study was set up to answer the research questions presented above. Figure 2 depicts the activity flow of the user study. Participants were requested to complete eight health scenarios using two search interfaces and two correctness settings: (SC) with a correct single-card, (SN) with an incorrect single-card, (MC) with multi-cards, including a correct card, and (MN) with multi-cards, where none of the cards are correct. A health card is correct when it matches the known diagnosis of the scenario (see Table 2).

The user study was performed in a usability laboratory with a PC equipped with eye tracking technology. To minimise fatigue bias, we rotated the eight scenarios and the four search interfaces using a Graeco-Latin square rotation, resulting in 32 scenario–search interface combinations. Participants were principally university students. The study
Table 1: Pre-task questionnaire items.

<table>
<thead>
<tr>
<th>Pre-task Questionnaire Items (options)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is the most probable health condition for the scenario? (open answer)</td>
</tr>
<tr>
<td>2. What would you do now? (1=Self-treat, 2=Contact an health professional, 3=Use an emergency service)</td>
</tr>
<tr>
<td>3. How confident are you with your answers? (1=Very not confident to 5=Very confident)</td>
</tr>
<tr>
<td>4. How many times have you searched for information about the topic of this scenario? (1=Never, 2=1-2 times, 3=3-4 times, 4=5 times)</td>
</tr>
<tr>
<td>5. How many times have you searched for information about the topic of this scenario? (1=Very not confident to 5=Very confident)</td>
</tr>
<tr>
<td>6. How interested are you to learn more about the topic of this scenario? (1=Very uninterested to 5=Very interested)</td>
</tr>
</tbody>
</table>

received Human Research Ethics Committee clearance (#2018002115). Next, we discuss each part of the user study.

Informed consent and demographic questionnaire

After consenting to participate, each participant was given a set of instructions presenting the elements of the interface and rules for the collection of evidence to answer the scenarios. Next, a demographic questionnaire collected information on the participant’s age group, highest level of education, education background, English proficiency\(\ast\) and the frequency of use of Web search engines. We used the responses to determine the participant’s eligibility.

Pre-task questionnaire

After completing the demographic questionnaire, participants completed pre-task questionnaires shown in Table 1 for all eight scenarios. We used the first three items to understand participants’ background knowledge for each health scenario and items 4 to 6 (adapted from Kelly et al.\(\ast\)) to understand the participant’s interest and background knowledge for each health scenario.

Search scenarios

We selected eight scenarios of the 45 standardised patient vignettes used in a survey of symptoms checkers.\(\ast\) The vignettes were compiled from various clinical sources such as education material for health professionals and a medical resource website. Each vignette contained age, gender, symptoms, correct diagnosis and correct category of triage urgency for a given condition. They include both common and uncommon diagnoses (based on prevalence) from three categories of triage urgency: requiring emergency care, requiring non-emergency care, and self-care appropriate. We ensured that each diagnosis in the eight selected scenarios had a matching Google health card.

Then, we created a topic description based on each vignette. A topic description contains all symptoms as reported by the patient in the vignette, excluding clinical observations (since in a real setting, the user would not have such information). We also replaced medical terms with layman terms, where appropriate (e.g., “rhinorrhea” was replaced with “runny nose” and “acetaminophen” was replaced with “paracetamol” as “paracetamol” is a more commonly known term in Australia than “acetaminophen”). Finally, we asked research students in our team lab (who have no medical background and had English as first language) to formulate a search query for each topic description. Table 2 reports the topic description, diagnosis, urgency category, and search query for each health scenario.

To complete each scenario, we asked participants to first make a diagnosis then copy and paste the condition mention — either from the snippets, linked documents, or from the health cards. This protocol allowed us to track where participants found the relevant diagnosis mention and evidence for making their health decision (i.e., they could have found it across different information objects, but they made their final decision based on the copied one). Second, we asked participants to select the urgency condition for the scenario: requires emergency care (e.g., calling 911 or immediately going to hospital), requires non-emergency care (e.g., contacting GP or nurse help line), or self-care appropriate (e.g., taking over the counter drug or home-remedy, resting, performing activities to mitigate the condition). Finally, we asked participants to rate their confidence of the responses (1=Very not confident to 5=Very confident), and the quality of the presented health card(s) with respect to three dimensions: relevancy, understandability and trustworthiness (1=[neg], 2=Partially [pos] 3=[pos]; where [neg], [pos] labels were contextualised to the items; the partial option was only shown when multi-cards were shown).

\(\ast\) We verified participants English proficiency by checking whether they: (1) spoke English as first language, or (2) achieved IELTS overall test score of at least 5.0 with a score of at least 4.5 in each of the four test components. These are the minimum English proficiency to work in Australia.
We ranked the identified health concepts (i.e., disease or syndrome) based on how many of the top 50 search results for extracting Unified Medical Language System (UMLS, version 2018AA) medical thesaurus entities from free-text. We identified disease or syndrome concepts in the title and the snippets of each search result using QuickUMLS in the interface (either from the results or from a health card), we presented them with the archived web page. We then noted by Jimmy et al.

To obtain the search results for each health scenario, we submitted the query to the Bing Web Search API thus, participants would be accustomed to the interface. The search interfaces contained three panes: left (not shown in Figure 1), middle, and right. The left pane displayed the topic, instructions, and tasks to be completed by participants. For the search interface with a single-card (Figure 1 left), the middle pane displayed the query string (disabled so a new query could not be entered) and the top ten search results (title, url, and snippet). The right pane of a single-card interface showed the health card. For the search interface with multi-cards (Figure 1 right), we merged the middle and the right panes to display the query string, the four health cards, and the top 10 search results. We designed the snippet list and the health cards following the Google search interface as it was the most popular search engine in Australia, i.e., where this study took place; however, we added multi-cards to accommodate the MultiSearch study participants. Since the MultiSearch study participants reported their overall experience in completing the tasks and their previous experiences in searching online for health information, with specific attention to the use of health cards. Due to space constraints, the analysis of the user experience questionnaire and exit questionnaire are out of the scope of this paper, and are left to future work.

**Search Interfaces**

The search interfaces (Figure 1) contained three panes: left (not shown in Figure 1), middle, and right. The left pane displayed the topic, instructions, and tasks to be completed by participants. For the search interface with a single-card (Figure 1 left), the middle pane displayed the query string (disabled so a new query could not be entered) and the top ten search results (title, url, and snippet). The right pane of a single-card interface showed the health card. For the search interface with multi-cards (Figure 1 right), we merged the middle and the right panes to display the query string, the four health cards, and the top 10 search results. We designed the snippet list and the health cards following the Google search interface as it was the most popular search engine in Australia, i.e., where this study took place; thus, participants would be accustomed to the interface.

**Search Results**

To obtain the search results for each health scenario, we submitted the query to the Bing Web Search API on February 2nd, 2019 and acquired the top 50 search results. To avoid problems with possible web pages and SERP updates, as noted by Jimmy et al., we archived all search results and source web pages. When a participant clicked on any link in the interface (either from the results or from a health card), we presented them with the archived web page. We then identified disease or syndrome concepts in the title and the snippets of each search result using QuickUMLS, a tool for extracting Unified Medical Language System (UMLS, version 2018AA) medical thesaurus entities from free-text. We ranked the identified health concepts (i.e., disease or syndrome) based on how many of the top 50 search results...
Table 3: Health cards for each scenario. [C] indicates the correct card.

<table>
<thead>
<tr>
<th>Id</th>
<th>Rank 1</th>
<th>Rank 2</th>
<th>Rank 3</th>
<th>Rank 4</th>
<th>Rank 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Septicemia</td>
<td>Stomach flu</td>
<td>Influenza</td>
<td>food poisoning</td>
<td>Appendicitis [C]</td>
</tr>
<tr>
<td>2</td>
<td>Migraine</td>
<td>Meningitis [C]</td>
<td>Encephalitis</td>
<td>Hypertensive disease</td>
<td>Pink eye</td>
</tr>
<tr>
<td>3</td>
<td>Hypertensive disease</td>
<td>Eclampsia</td>
<td>Type 2 diabetes</td>
<td>Venous ulcer</td>
<td>Deep Vein Thrombosis [C]</td>
</tr>
<tr>
<td>4</td>
<td>Anorexia</td>
<td>Common Cold</td>
<td>Roseola</td>
<td>Sinusitis</td>
<td>Otitis Media [C]</td>
</tr>
<tr>
<td>5</td>
<td>Sinusitis [C]</td>
<td>Upper respiratory infection</td>
<td>Chronic obstructive pulmonary disease</td>
<td>Perianal abscess</td>
<td>Common Cold</td>
</tr>
<tr>
<td>6</td>
<td>Acute Bronchitis</td>
<td>Pneumonia</td>
<td>bronchiolitis</td>
<td>Cheilitis</td>
<td>Chronic obstructive pulmonary disease [C]</td>
</tr>
<tr>
<td>7</td>
<td>Upper respiratory infection</td>
<td>Bronchitis [C]</td>
<td>Chronic obstructive pulmonary disease</td>
<td>Seasonal allergies</td>
<td>Common Cold</td>
</tr>
<tr>
<td>8</td>
<td>Blepharitis</td>
<td>Pink eye</td>
<td>Stye [C]</td>
<td>Chapped lips</td>
<td>Chalazion</td>
</tr>
</tbody>
</table>

contained each concept and we kept the five most frequent health concepts. We ensured the selected five concepts contained four incorrect health concepts and one correct health concept (there could not be more than one correct health concept for a scenario). If the top five concepts were all incorrect, we exchanged the lowest ranked with the correct concept. This was so that in the multi-cards interface we could display either four incorrect or three incorrect and one correct health cards. We only considered health concepts that matched a Google health card. Table 3 lists the five cards for each health scenario. Finally, we selected the two search results with the highest rank for each health concept, to make up in total 10 search results for each health scenario, to display in the SERP. The 10 search results were ordered based on ranks from Bing.

Health Cards
Health cards were acquired from the Google search engine based on Table 3. Each health card contained a title, aliases (i.e., “also called”), if any, an image, a summary tab (i.e., about), a symptoms tab, and a treatments tab. Each tab contained a URL that links to the source information for the health card. In the acquired Google health cards, the word “critical” in bold would appear in the bottom part of the card if the associated condition requires emergency medical attention (e.g., “Appendicitis”); we maintained this feature. For health cards that had no image from the Google health cards, we obtained an image from other medical web pages that discuss the same condition. We fixed the health cards height to 600px and summarised the health cards content to ensure that all information fit the height setting. This was done to provide a similar look & feel for all search interfaces in the study.

Capturing Interaction Data
Throughout the user study, we captured participants interactions with the search interfaces using the Big Brother logging service. Big Brother records mouse movements (including anchored to <div> containers, e.g., enter and leave the container), clicks, scroll, page loading (start and end), cut/copy/paste, and window’s scroll position (mainly to align and validate eye-tracking data).

We used the Tobii Pro Spectrum eye tracker to acquire eye gaze data, set to operate at the frequency of 300Hz. The eye tracker was connected to a monitor with a resolution of 1920 x 1080 pixels. The eye tracker was calibrated for each participant at the start of the study using the method described by Blignaut. We implemented the velocity-threshold identification algorithm to identify fixation points. We set the velocity radius threshold to 70 pixels following the size of the eye gazing point visualisation from the Tobii Pro Eye Tracker Manager. We set the minimum fixation duration threshold to 100ms as suggested by Blignau. Then, we mapped the fixation points to three areas-of-Interest (AOIs): scenario description (left pane), list of snippets, and health cards.

The eye gaze data was used to determine whether participants noticed the health cards, and how much time they spent on the health card, compared to the rest of the SERP or actual result web pages. Other analyses of the collected eye

‡https://github.com/hscells/bigbro
tracking data was regarded as being out of scope of this paper, and left to future work.

**Participants**
The study was advertised widely through The University of Queensland, a large public university in Australia, as well as through Facebook groups mainly tailored to students and alumni of this university. We did not enforce participants to be university students or affiliates (but we excluded research staff), and we allowed any member of the public to take part in the study. Nevertheless, the majority of the participants were university students. The following eligibility criteria for participation in the study were set and enforced: aged 18 years or above, no specific prior medical studies, experienced with using a web search engine on a daily basis, and proficient English readers and writers. Participants were told that the study would last approximately one hour and were given a AUD$15 gift card for their participation.

**Experiment Results**
From 64 participants, each performing eight scenarios, we collected 512 interaction data points. This gave us enough power for statistical analysis (power > 0.90), statistical tests using unpaired t-test adjusted with Bonferroni correction due to multiple comparisons in an analysis. Each of the 16 sequences of scenarios-search interface pairs was performed by three participants. Participants comprised 38 females and 26 males: 31 between 18-24 y.o., 23 between 25-34 y.o., 9 between 35-44 y.o. and 1 between 45-54. Their education background was: 20 Engineering/IT, 13 Business/Economics, 13 Science, 5 Psychology, 6 peripherally health related majors with no medical or diagnostic experience (e.g., personal trainer), 7 Humanities/Social Science. We confirmed that no participant had experience in the medical field and specifically in diagnosing health conditions. The highest level of completed education was: 13 high school, 13 diploma, 18 bachelor degree, 5 graduate diploma, and 15 postgraduate degree.

Participants considered the eight scenarios as interesting and unfamiliar. Overall, the participants’ average level of interest in the scenarios was moderately high (3.9 with, 1=very not interesting to 5=very interesting). Further, participants had no to little knowledge of the scenarios (M = 1.88) and never or rarely searched for information related to the scenarios (M = 1.42). We found very weak to no correlation between the participants’ level of interest, prior knowledge and prior search for the scenario on all six measurements; thus these had no systematic effect on results.

Then, we investigated whether fatigue may have had a systematic effect on results. We did this by correlating the sequence of scenarios and the results from the six measurements used in RQ1. We found that there was a weak negative correlation between the scenario sequence and duration taken to complete a scenario (corr=-0.21): this may have been due to fatigue or acquired familiarity with the search tasks and interface. Further, we found very weak to no correlation between the scenario sequence and the other five measurements: health card usage rate (corr=0.02), correct diagnosis rate (corr=-0.05), correct urgency score (corr=0.06), number of page read (corr=0.04), good abandonment rate (corr=0.00), and confidence rate (corr=-0.13). This indicates that the results are comparable across scenario sequences. Next, we investigated the impact of presenting single and multi-cards on decision making behaviour.

The multi-cards interface was found to drive users’ attention toward the health cards (as opposed to the search results). When the multi-cards were shown, most participants spent more time on health cards (M = 82%) than on snippets (M = 18%). However, when only a single-card was shown, participants spent equal amounts of time on snippets and on health cards (M = 51% vs. M = 49%). These findings were observed by measuring which AOI (i.e., snippets vs. health cards) participants paid attention to when health cards were displayed. (We removed eye tracker data associated to other display areas, e.g., the instructions pane.) Since the time taken by each session varied, we normalised durations, and present results with respect to the progress in the session. Figure 2 reports the percentage of participants that paid attention to each AOI throughout the session. These findings are understandable since when the single-card is shown, there is more information to process in snippets and the display area containing the snippets is larger. On the other hand, the multi-cards occupy most of the initial display and only the top three snippets are visible without scrolling. Participants may have also found most of the information they required among the multi-cards.

Participants often considered health cards earlier in their session: There is a moderate to strong negative correlation between attention on health cards (according to eye tracking) and time point in the session when both single-card (corr=-0.50) and multi-cards (corr=-0.90) were shown. Interestingly, regardless of the declining attention overtime, we found that the attention on health cards increased at the end of the session for both interfaces. We speculate that...
although health cards were prioritised at the beginning of the session, participants may have felt the health cards did not contain enough information to make the final decision, and went on examining snippets throughout the SERP. Participants may have felt that information from the snippets was still not sufficient to complete the scenario, hence, they re-considered the health cards at the end of the session.

Results for RQ1: How do single and multiple health cards influence CHS users when making health decisions?

Regardless of the search interfaces, participants perceived health cards as relevant ($M = 0.8203$), trustworthy ($M = 2.5762$) and easy to understand ($M = 2.703$). We measured health cards’ relevance as binary where 0=not relevant and 1=relevant (Note that a health card that was considered as relevant, was not necessarily considered as correct). We speculate that participants may have considered an incorrect health card as relevant since it helped participants to rule out a condition. Trustworthiness and understandability were measured on a 1 to 3 scale (3 being most trustworthy/understandable). Next, we contrast the influence of single and multiple health cards on CHS decision making based on results on the key metrics reported in Table 4.

Multi-cards were preferred as a source of information. This is evident from the probability that the information in health cards was selected for making the health decisions. When multi-cards were presented, in fact, most participants preferred to select information from one of the health cards to complete the scenario (68.75% of 256 scenarios), while when the traditional single-card was presented, the organic search results were preferred more than the health card (only 37.11% of 256 scenario show information selected from the health card). This was regardless of the correctness of the health cards, and differences are strongly statistically significant.

In terms of diagnosis correctness, the single-card was more effective than the multi-cards in leading participants to identify the correct diagnosis (37.89% vs. 24.61%): these differences are strongly statistically significant. Similarly, for the level of urgency correctness, we also found that the average correctness score of the submitted urgency was

<table>
<thead>
<tr>
<th>Measurements</th>
<th>Num. Cards Shown</th>
<th>Correct Card Present</th>
<th>Search Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single$^a$</td>
<td>Multi$^b$</td>
<td>Yes$^c$</td>
</tr>
<tr>
<td>1. Health cards usage</td>
<td>0.3711$^{B}$</td>
<td>0.6875$^{A}$</td>
<td>0.5664</td>
</tr>
<tr>
<td>2.A. Correct diagnosis</td>
<td>0.3789$^{B}$</td>
<td>0.2461$^{A}$</td>
<td>0.4922$^{B}$</td>
</tr>
<tr>
<td>2.B. Urgency score</td>
<td>0.543$^{b}$</td>
<td>0.4453$^{a}$</td>
<td>0.543$^{b}$</td>
</tr>
<tr>
<td>3. Duration (sec.)</td>
<td>157</td>
<td>158</td>
<td>150</td>
</tr>
<tr>
<td>4. Num. of page opened</td>
<td>2.4844$^{B}$</td>
<td>1.3438$^{A}$</td>
<td>1.7812</td>
</tr>
<tr>
<td>5. Good abandonment</td>
<td>0.2344$^{B}$</td>
<td>0.4727$^{A}$</td>
<td>0.3789</td>
</tr>
</tbody>
</table>
significantly higher when a single-card was shown (0.543) than when multi-cards were shown (0.4453). To determine the correctness score of the submitted urgency, we computed \( \text{score} = 1 - (|\Delta \text{urgency}| \cdot p) \) where \( \Delta \text{urgency} = \text{correct urgency} - \text{submitted urgency} \) and the possible level of urgency are: 1= self-treat, 2= contact a health professional, and 3= use an emergency service. Lastly, \( p \) models the penalty for an incorrect urgency decision. We set the penalty for incorrect decisions so as to greatly penalise urgency decisions that put the well-being of the person at risk (e.g., decided to self-treat when the correct urgency was to use an emergency service). For this study, we set \( p = 1 \) when \( \Delta \text{urgency} \geq 0 \) and \( p = 0.5 \) when \( \Delta \text{urgency} < 0 \).

The multi-cards interface, on average, lead participants to identify lower levels of urgency (not the correctness score). In fact, a significantly higher level of urgency was recorded when a single-card was shown (\( M = 1.9258 \)) than when multi-cards were shown (\( M = 1.7813 \)). Yet, the mean of the submitted level of urgency from both settings were lower than the mean of the correct level of urgency (2.125). This suggests that participants were more likely to make incorrect “what to do” decisions that could have put the person in the scenario at risk.

The interface type did not influence the session duration (no significant difference, \( M = 158 \text{seconds} \)); however the multi-cards required less efforts (clicking and browsing web pages) from participants than the single-card interface (this difference is statistically significant). In fact, regardless of the presence of a correct card, users opened significantly less result pages when provided with multi-cards. Furthermore, results also show that multi-cards were more likely to lead participants to good abandonment\( ^2 \) compared to the single-card (difference is statistically significant). Note that good abandonment occurs when a scenario is completed without clicking on links from the SERP. Interestingly, although not statistically significant, we found that the likelihood of participants submitting a correct diagnosis was higher when good abandonment occurred (35.36% of 181 scenarios) than when clicking on links (29%). Of the 181 good abandonment occurrences, 83.43% of the decisions were made by selecting information from health cards.

Finally, we found that there were no significance differences in the level of confidence in the decisions made across all search interfaces. Interestingly, in line with Zeng et al.'s findings\( ^1 \), we found no correlation between the level of confidence in the user decisions and the user decisions’ correctness (both diagnosis correctness and urgency score). The majority of the diagnosis decisions (66%) submitted with (very) confidence were incorrect.

Of the six measurements, only the diagnosis decisions’ correctness suggests that the multi-cards are significantly less useful than the single-cards for assisting decision making in CHS. In fact, the highest probability of making a correct diagnosis (57.81%) and the highest average urgency correctness score (0.5977) were found when single correct cards were shown. Nevertheless, determining the single-card that is relevant to a user’s query is not trivial for a search system. A search system may identify the single-card to show by relying on the top search result; alternatively it could rely on ranking cards according to the popularity of the underlying concept within the search results. However, if a system based on a single card interface was to follow any of these two methods, then it would display the correct card for only 1 of the 8 queries in Table\( ^3 \) (thus, probability of single correct card shown: \( P(SC) = 0.125 \)). This is unlike for a multi-cards interface, for which the probability of correct card shown: \( P(MC) = 0.5 \). Overall, the probabilities that a correct diagnosis is made depending on the two interfaces can be compared:

\[
P(\text{correct}|S) \geq P(\text{correct}|M) \Rightarrow \\
P(SC) \cdot P(\text{correct}|SC) + P(SN) \cdot P(\text{correct}|SN) \leq P(MC) \cdot P(\text{correct}|MC) + P(MN) \cdot P(\text{correct}|MN) \Rightarrow \\
0.125 \cdot 0.5781 + 0.875 \cdot 0.1797 \leq 0.5 \cdot 0.4062 + 0.5 \cdot 0.0859 \Rightarrow 0.2295 \leq 0.2461 \tag{1}
\]

The result above suggests that, if the probability of the system showing a correct health card was accounted for, then the multi-cards interface is more likely to lead to correct diagnoses.

**Results for RQ2: How accurate are CHS users in appraising the correctness of health cards?**

Next, we only focus on the MC interface to investigate the users’ ability in identifying the correct health card. We tied the correctness of the appraisal of an information object (cards, snippets) to the diagnosis decision that participants made based on the information object, which is identified by the source of information measurement. That is, a correct appraisal of a health card, for example, occurs when a correct decision is made and information from that health card is selected in their answer.

We found that the majority of participants were not able to assess the correctness of a health card within the MC interface. In fact, of all decisions taken using this interface, 39.1% were taken based on incorrect health cards (and
which lead to incorrect diagnoses), while 33.6% were taken based on correct health cards. This situation is however better than when users attempted to assess snippets. In this case, in fact, the majority of decisions taken using snippets lead to incorrect diagnoses (20.3%), while only 7.0% were correct. Remember that in every SERP, a minimum of two out of 10 snippets contained the correct diagnosis.

In summary, the appraisal of health cards by users is still a challenge, but health cards are more likely to be correctly appraised than search results snippets.

Discussion

Diagnosis correctness was found to be dependent on the correctness of the presented health cards, in line with findings of Pogacar et al. 

Regardless of the interface (single vs. multi-cards), the correctness of the diagnosis was statistically significantly higher when a correct card was presented (0.4922) than when no correct card was presented (0.1328). The highest level of correctness was found when only the correct health card was presented.

Participants consistently underestimated the necessary level of urgency (e.g., decide to self-treat when they should have gone to the Emergency); this finding is independent of the correctness of the diagnosis. Figure 4 shows that, before searching, 51% of 512 tasks have been assigned the wrong level of urgency. Of these, 57% (43%) were underestimated (overestimated). In fact, search activities lead to a strongly significantly increase in number of correct diagnoses (4% correct diagnoses prior to search VS. 38% (single-card) and 25% (multi-cards) post search). However, the number of correctly estimated urgency levels are comparable prior (49%) and post search (50% for single-card; 44% for multi-cards). Of the incorrectly estimated levels of urgency post search, 67% (single-card) and 76% (multi-card) of the times these were underestimated. When considering decisions for which a correct diagnosis was identified, 43% of these had an incorrect urgency level (73% of which were underestimated).

Conclusions

We investigated the influence of health cards to assist decision making in consumer health search. Specifically, we presented eight health scenarios and asked 64 participants to make two decisions: determining the portrait health condition and determining the follow up action that should be taken (level of urgency). The experiment was conducted in a laboratory using two search interfaces: one with a traditional single card and one with a novel multi-cards design. Regardless of the search interfaces, health cards were perceived as relevant, trustworthy and easy to understand.

We found that showing a correct single card leads to the highest probability of identifying a correct diagnosis and level of urgency. However, determining the correct single card that is relevant to a user’s query is not trivial for a search system, and errors are made when an incorrect card is presented instead. If the probability of the system showing a correct health card is accounted for, then the multi-cards interface (which is more likely to show a correct card) leads to a higher number of correct diagnosis decisions than the single-card interface. We also found that participants consistently underestimated urgency across settings: this is of particular concerns because, while an incorrect diagnosis can be rectified by a clinician, this would not happen if users do not seek medical attention.

Furthermore, we found that the multi-cards interface enables users to make health decisions with significantly less effort than the the single card interface. When multi-cards were shown, in fact, participants clicked significantly less links, and completed significantly more scenarios without clicking on links from the SERP (i.e., good abandonment), than when they were presented with a single-card.

The multi-cards were the most preferred source of information, compared to the traditional single cards and the search results snippets. From the eye-gazing data, in fact, we found that when a single-card was shown, participants spent a
comparable amount of time considering the health cards and the snippets. On the other hand, when multi-cards were shown, participants spent significantly more time considering the health cards, compared to the snippets.

Finally, we found that the appraisal of search results is still an issue, even within the multi-cards interface (MC). In fact, only 33.6% of decisions taken in this context and using the health cards lead to a correct diagnosis. However, this was better than when using snippets only (7.0%).

Our findings suggest that, from a search engine designer’s perspective, the provision of health cards (in addition to organic search results) should generally be preferred to the display of search results only. However, care should be taken in selecting the health cards to display, as their correctness influences the correctness of health decisions (both in terms of diagnosis and urgency). To address this aspect, in future work we plan to investigate and advance the effectiveness of current entity-oriented retrieval methods to identify and rank health cards.

A limitation of our study is that participants were recruited from a university campus and this may not be representative of the real diversity of users (e.g., our sample of users was biased towards young people). In future work we plan to overcome this limitation by considering other means of participant recruitment and by increasing the sample size.

Acknowledgments. Jimmy is sponsored by the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP) (20151022014644). Dr. Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579) and a Google Faculty Award; both these grants funded this study.

References
Assessing the Validity of a *priori* Patient-Trial Generalizability Score using Real-world Data from a Large Clinical Data Research Network: A Colorectal Cancer Clinical Trial Case Study

Qian Li, MS¹, Zhe He, PhD²#, Yi Guo, PhD¹#, Hansi Zhang, MS¹, Thomas J George Jr, MD, FACP¹, William Hogan, MD, MS¹, Neil Charness, PhD², Jiang Bian, PhD¹*

¹University of Florida, Gainesville, FL, USA; ²Florida State University, Tallahassee, FL, USA;

Abstract

Existing trials had not taken enough consideration of their population representativeness, which can lower the effectiveness when the treatment is applied in real-world clinical practice. We analyzed the eligibility criteria of Bevacizumab colorectal cancer treatment trials, assessed their *a priori* generalizability, and examined how it affects patient outcomes when applied in real-world clinical settings. To do so, we extracted patient-level data from a large collection of electronic health records (EHRs) from the OneFlorida consortium. We built a zero-inflated negative binomial model using a composite patient-trial generalizability (cPTG) score to predict patients’ clinical outcomes (i.e., number of serious adverse events, [SAEs]). Our study results provide a body of evidence that 1) the cPTG scores can predict patient outcomes; and 2) patients who are more similar to the study population in the trials that were used to develop the treatment will have a significantly lower possibility to experience serious adverse events.

Introduction

Clinical studies (trials) are essential in evidence-based medicine.¹ Clinical trials, however, are often conducted under idealized and rigorously controlled conditions to improve their internal validity and success rates; but such conditions, paradoxically, may compromise their external validity (i.e., trial results’ generalizability to the real-world target populations).² These idealized conditions are sometimes exaggerated and reflected as overly restrictive eligibility criteria. The generalizability and study population representativeness have long been major concerns.²,³ Certain population subgroups, such as older adults,⁴,⁵ are often underrepresented due to unjustified exclusion criteria, especially in cancer studies.⁶⁻¹⁰ The underrepresentation of these population subgroups can lead to low trial generalizability, and subsequently, reduce treatment effectiveness and increase the likelihood of adverse outcomes in these population subgroups when the treatments are moved into real-world clinical practice. As a consequence, some approved drugs have been withdrawn from the market after serious adverse drug reactions were observed on population subgroups excluded from the original trials.¹¹

In clinical trials, the target population represents the patients to whom the results of the clinical trials are intended to be applied. The study population represents the patients being sought as defined in the clinical trial eligibility criteria (Figure 1). To ensure patient safety and demonstrate treatment efficacy, eligibility criteria are often restrictive, thus representing a constrained subset of the target population. Further, study samples are the enrolled participants of a trial. Even though study participants are screened based on eligibility criteria, due to real-world constraints such as trial awareness, the study sample may not adequately represent the study population defined by the eligibility criteria. Further, it is worth noting that the notions of generalizability and population representativeness are related but distinct. Population representativeness measures the study population’s coverage of real-world patients, with respect to study traits (e.g., age, vitals, and labs), often defined by eligibility criteria. On the other hand, generalizability is the portability of the causal effects of an intervention to real-world settings. Besides population representativeness, other factors also affect studies’ generalizability, such as variation in patients across different clinical settings, discrepancies in conditions under which a trial was conducted, and incomplete reporting.¹²,¹³ Nevertheless, population representativeness of the study population is one of the determining factors for its generalizability.

---

# Contributed equally, co-first authors
* Corresponding: Jiang Bian; bianjiang@ufl.edu

Figure 1. Populations in clinical trials.
There are two major types of approaches to assessing a study’s population representativeness: 1) the \textit{a priori} generalizability is the representativeness of eligible participants (study population) to the target population; 2) the \textit{a posteriori} generalizability is the representativeness of enrolled participants (study sample) to the target population. The \textit{a priori} generalizability can also be called as the eligibility-driven generalizability, whereas the \textit{a posteriori} generalizability can be called as the sample-driven generalizability.\textsuperscript{14} However, the \textit{a posteriori} generalizability is an artifact after the fact (i.e., can only be done when a trial is concluded and complete trial data are collected), while the \textit{a priori} generalizability is affected by the trial’s eligibility criteria that are modifiable during the study design phase.

In 2014, an \textit{a priori} generalizability score “Generalizability Index for Study Traits” (GIST 1.0; we use 1.0 to differentiate these from the GIST 2.0 metrics introduced below) was introduced to quantify the population representativeness using eligibility criteria one at a time and real-world patient data.\textsuperscript{15} The GIST 1.0 score characterizes the proportion of patients that would be potentially eligible across trials with the same trait over the target population. GIST 1.0 was validated using a simulated target population.\textsuperscript{16} As some criteria are correlated, GIST was extended to mGIST 1.0 with joint use of multiple eligibility criteria.\textsuperscript{17} Later, GIST 2.0\textsuperscript{18} was developed to consider both 1) the dependencies across multiple criteria in a study, and 2) the significance of individual traits across different diseases (e.g., HbA1C is more important in type 2 diabetes than it is in chronic kidney disease). GIST 2.0 has two components: sGIST (w.r.t. one criterion) and mGIST (w.r.t. multiple criteria) corresponding to the original GIST 1.0 and mGIST 1.0, respectively. GIST scores are between 0 and 1, where a higher score indicates a greater population representativeness. From now on, we will use sGIST and mGIST to refer to the GIST 2.0 metrics.

Nevertheless, to be able to rationalize adjustments of eligibility criteria towards ultimately better generalizability early on, it is important to identify the relationships among the \textit{a priori} generalizability and the actual outcomes of the interventions/treatments in real-world settings. Treatment outcomes can be measured in several different ways: 1) improvement in clinical outcomes (e.g., better lab results), 2) less adverse events (AEs); and ultimately 3) longer survival and better quality-of-life (QoL). Adverse event is an important measure of treatment safety during clinical trials. Previously, Sen et al. found that GIST 2.0 score was significantly correlated with the number of AEs (i.e., the lower the GIST 2.0 score the higher the number of AEs) based on data from 16 sepsis trial results sections in ClinicalTrials.gov.\textsuperscript{14} Nevertheless, their study is limited as it does not provide any evidence on whether the \textit{a priori} generalizability of the trials had any impact on the clinical outcomes (i.e., the number of AEs) when the treatment is applied to the target patients in real-world clinical practices.

On the other hand, the wide adoption of electronic health record (EHR) systems and the proliferation of clinical data warehouses with rich real-world patient datasets offer unique opportunities to address these studies. The U.S. Food and Drug Administration (FDA) recently coined the terms real-world evidence (RWE) and real-world data (RWD) as “data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than traditional clinical trials.”\textsuperscript{19,20} RWD can come from various sources including EHRs, claims and billing activities as well as patient-generated data. These data will play an increasingly important role in health care and regulatory decisions.

In this study, we aim to fill this important knowledge gap exploring the hypotheses that 1) the \textit{a priori} generalizability of treatment trials is correlated with the clinical outcomes of the treatment (i.e., the number of AEs) and 2) patients who are eligible (i.e., based on the trials’ eligibility criteria and patient characteristics in their EHRs) for the original trials used to develop the treatment will have better clinical outcomes than those who are not, when the treatment is applied on both eligible/ineligible patients in real-world clinical settings, using RWD—a large collection of linked EHRs and claims—from a large clinical data research network. Our current study focuses on Bevacizumab (under the trade name Avastin)—a first of its kind of monoclonal antibody as a tumor-starving (anti-angiogenic) therapy—approved by the FDA in 2006 for the treatment of metastatic colorectal cancer.\textsuperscript{21}

\section*{Methods}

\subsubsection*{Data sources}

\textbf{Bevacizumab (Avastin) clinical trials and trial eligibility criteria.} We obtained free-text eligibility criteria from ClinicalTrials.gov—a registry maintained by the National Library of Medicine (NLM) in the United States. As of March 2019, over 299,335 studies across all 50 states in the US as well as in 208 countries are registered on ClinicalTrials.gov. Study information in ClinicalTrials.gov is semi-structured: study descriptors such as study phase, intervention type, and locations are stored in structured fields while eligibility criteria are largely free-text. Through ClinicalTrials.gov, we found 57 Bevacizumab trials that met our inclusion criteria: 1) the trial was conducted in the US; 2) the primary purpose of the trial is the development of the treatment agent; and 3) excluding post-market observational studies (e.g., comparative effectiveness studies, and essentially all Phase IV studies).
Real-world patient data from the OneFlorida. We obtained individual-level patient data from the OneFlorida Clinical Research Consortium (OneFlorida CRC), one of the 13 Clinical Data Research Networks (CDRNs) contributing to the national Patient-Centered Clinical Research Network (PCORnet) previously funded by the Patient-Centered Outcomes Research Institute (PCORI). PCORnet is now supported by the newly incorporated People Centered Research Foundation (PCRF), a nonprofit formed by PCORnet investigators with significant additional infrastructure-building funds from PCORI. The OneFlorida data repository integrated various data sources from contributing organizations in the OneFlorida CRC currently including 10 healthcare organizations: 1) two academic health centers (i.e., University of Florida Health, UFHealth and University of Miami Health System, UHealth), 2) seven healthcare systems including Tallahassee Memorial Healthcare (TMH affiliated with Florida State University), Orlando Health (ORH), Adventist Health (AH, formerly known as Florida Hospital), Nicklaus Children’s Hospital (NCH, formerly known as Miami Children’s Hospital), Bond Community Health (BCH), Capital Health Plan (CHP), and Health Choice Network, (HCN), and 3) CommunityHealth IT—a rural health network in Florida. In addition, OneFlorida also obtained claims data from the Florida Medicaid (FLM) program. As a network, the OneFlorida CRC provides care for approximately 48% of Floridians through 4,100 physicians, 914 clinical practices, and 22 hospitals with a catchment area covering all 67 Florida counties. Most HCOs in OneFlorida contributed EHRs, while CHP and FLM contributed claims data. We linked patients across the different EHRs and claims data sources using a validated privacy-preserving record linkage method. OneFlorida contains only a limited data set under the Health Insurance Portability and Accountability Act (HIPAA) and follows the PCORnet Common Data Model (CDM) v4.1 including patient demographics, enrollment status, vital signs, conditions, encounters, diagnoses, procedures, prescribing (i.e., provider orders for medications), dispensing (i.e., outpatient pharmacy dispensing), and lab results. The scale of the data is ever growing with a collection of longitudinal and robust patient-level records of ~15 million Floridians and over 463 million encounters, 917.6 million diagnoses, 1 billion prescribing records, and 1.17 billion procedures as of December 2018. Since our goal is to evaluate the generalizability of Bevacizumab trials as a colorectal cancer treatment, we extracted colorectal patients from OneFlorida as the target population using ICD-9/10-CM codes (i.e., ICD-9: 153.*, 154.*, 159.0; ICD-10: C18.*, C19.*, C20.*, C26.0). We identified 39,776 unique colorectal patients and extracted their data from OneFlorida.

Figure 2. Overall study design and selection of study populations.

**Overall study design**

There are 3 different populations in our study: 1) population A: OneFlorida colorectal cancer patients who were treated with Bevacizumab; 2) population B: patients who were treated with Bevacizumab and eligible for the original trials
Our hypothesis is that population A and population B will have better outcomes (i.e., less AEs) compared with population C. To do so, we devised a composite patient-trial generalizability (cPTG) score based on GIST that considers both trial generalizability and patient eligibility of the individual trials.

Our analysis consists of 7 steps: 1) analyzing Bevacizumab colorectal cancer treatment trials extracted from ClinicalTrials.gov to determine the computability of each eligibility criterion and constructing queries to extract study traits corresponding to each trial eligibility criterion based on the OneFlorida data; 2) identifying the different populations of interest (i.e., populations B and C as described above); 3) calculating each trial’s GIST score based on the computable eligibility criteria of each trial; 4) determining the eligibility (i.e., a binary variable) of each patient for each Bevacizumab trial of interest; 5) calculating the composite patient-trial generalizability score for each patient; 6) identifying each patient’s clinical outcome (i.e., the number of serious AEs) of being treated with Bevacizumab; and 7) comparing the difference in the numbers of serious AEs between population B and population C and examining the relationships between the cPTG score and the number of serious AEs.

### Defining populations of interest and serious adverse events (SAEs) related to using Bevacizumab

We used International Classification of Diseases, Ninth Revision/Tenth Revision, Clinical Modification (ICD-9/10-CM) codes to identify patients who were diagnosed with colorectal cancer in the OneFlorida data. Within the colorectal cancer patients, we then used the Healthcare Common Procedure Coding System (HCPCS) codes (e.g., C9257, J9035) combined with Bevacizumab’s RxNORM (e.g., 337521) codes to identify the administration of Bevacizumab. To identify SAEs in patients treated with Bevacizumab, we first reviewed the FDA approved drug label of Bevacizumab using the DailyMed database maintained by NLM. Then for each SAE, we identify the corresponding ICD-9/10-CM codes, as shown in Table 1.

### Table 1. ICD-9/10 codes to define serious adverse events for Bevacizumab based on its drug label.

<table>
<thead>
<tr>
<th>Serious Adverse Event</th>
<th>ICD-9-CM</th>
<th>ICD-10-CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptomatic postprocedural ovarian failure</td>
<td>256.2</td>
<td>E89.40</td>
</tr>
<tr>
<td>Systolic (congestive) heart failure</td>
<td>428.2</td>
<td>I50.2</td>
</tr>
<tr>
<td>Diastolic (congestive) heart failure</td>
<td>428.3</td>
<td>I50.3</td>
</tr>
<tr>
<td>Combined systolic and diastolic heart failure</td>
<td>428.4</td>
<td>I50.4</td>
</tr>
<tr>
<td>Cerebral Hemorrhage</td>
<td>431</td>
<td>I60</td>
</tr>
<tr>
<td>Intracranial Hemorrhage</td>
<td>432.9</td>
<td>I62.9</td>
</tr>
<tr>
<td>Fistula of intestine</td>
<td>569.81</td>
<td>K63.2</td>
</tr>
<tr>
<td>Perforation of intestine</td>
<td>569.83</td>
<td>K63.1</td>
</tr>
<tr>
<td>Gastrointestinal hemorrhage</td>
<td>578</td>
<td>K92.2</td>
</tr>
<tr>
<td>Hematemesis</td>
<td>578.0</td>
<td>K92.0</td>
</tr>
<tr>
<td>Vaginal bleeding</td>
<td>626.8</td>
<td>N93.9</td>
</tr>
<tr>
<td>Epistaxis</td>
<td>784.7</td>
<td>R04.0</td>
</tr>
<tr>
<td>Hemoptyisis</td>
<td>786.3</td>
<td>R04.2</td>
</tr>
<tr>
<td>Proteinuria</td>
<td>791.0</td>
<td>R80</td>
</tr>
</tbody>
</table>

As shown in Figure 3, to count as a SAE related to Bevacizumab, the SAE diagnosis code has to occur 1) after the first Bevacizumab procedure, but 2) within 180 days after the last Bevacizumab procedure. Note that the first Bevacizumab procedure has to occur after the diagnosis of colorectal cancer. Using these criteria, we counted the total number of SAEs for each patient.

### A new composite patient-trial generalizability (cPTG) score based on GIST

Figure 3. Selection window for serious adverse events related to treating colorectal cancer with Bevacizumab.
The original GIST metric quantifies the \textit{a priori} generalizability of clinical trials with respect to selected quantitative eligibility criteria that specify a permissible value range (e.g., HbA1c $> 7\%$), one at a time.\textsuperscript{15} The GIST score ranges from 0 to 1, with 0 being not generalizable and 1 being perfectly generalizable. In essence, it characterizes the proportion of patients potentially eligible across trials. The extension of GIST, mGIST,\textsuperscript{17} can quantify the population representativeness with joint use of multiple criteria of interest. Both GIST and mGIST focus on the generalizability assessment at the disease domain level (i.e., assessing the generalizability of trials targeting the same disease). GIST 2.0 was then introduced as a scalable multivariate metric for quantifying the population representativeness of individual clinical trials by explicitly modeling the dependencies among all eligibility criteria.\textsuperscript{18} The original implementation of GIST 2.0 was in Matlab. We implemented GIST 2.0 in Python and made it available in a public Github repository (i.e., \url{https://github.com/Andeeli/GIST.git}). The calculation of sGIST and mGIST is trivial and details can be found in the original GIST 2.0 publication.\textsuperscript{18} One key input is to define the target population (i.e., the real-world patient population that the treatment is intended to be applied on). For our study, we defined patients with colorectal cancer and treated with Bevacizumab (i.e., population A) as our target population. The outputs of GIST 2.0 are single-trait GIST score (sGIST) for each trait and one multiple-trait GIST score (mGIST) for the trial.

As the mGIST score is a trial-level variable, we also need to determine whether a patient is thought to be eligible of a trial based on its eligibility criteria and corresponding patient traits as defined in their EHRs (i.e., based on OneFlorida data in our case). The process is straightforward. For example, “Platelet count at least 1,500/mm$^3$” is an inclusion criterion in trial NCT00025337. The lab results for platelet tests are coded using Logical Observation Identifiers Names and Codes (LOINC) (i.e., 26515-7, 777-3, and 778-1) in OneFlorida. Based on these LOINC codes, we queried patients’ platelet test results and transformed the measurement unit to mm$/^3$ if needed. If a patient’s platelet lab result is larger or equal to 1,500/mm$^3$, we will consider the patient as met this particular inclusion criterion. We then determined the patient’s eligibility (i.e., a binary variable) of the trial, when the patient met all the inclusion criteria while did not meet any of the exclusion criteria. Note that not all eligibility criteria are computable as the needed data elements may not exist in their EHRs (e.g., “Fertile patients must use effective contraception”). We did not use these non-computable criteria when determining a patient’s eligibility.

In order to consider both the patient-level eligibility and trial-level \textit{a priori} generalizability, we propose a composite patient-trial generalizability (cPTG) score. To calculate cPTG score, we first calculated the mGIST score (ranging from 0 to 1) for each trial. Then, we used patients’ traits data (from OneFlorida) to create an index (i.e., 0 as not eligible or 1 as eligible) to indicate whether a patient is eligible for the specific trial or not. We then took the average of the dot product of the vector of mGIST and the vector of patient’s eligibility as follows:

$$cPTG_{i} = \frac{1}{K} \sum_{j=1}^{K} e_{ij} g_{j}$$

where $e_{ij}$ is the eligibility (0 or 1) for patient $i$ of trial $j$, and $g_{j}$ is the generalizability score (i.e., mGIST) of trial $j$, for each patient $i$ (i.e., $i=1,2,...,N$) and each trial $j$ (i.e., $j=1,2,...,K$). The cPTG score ranges from 0 to 1. Intuitively, a higher cPTG score for a patient means the patient is eligible for more trials in the set of trials of interest and those trials have higher \textit{a priori} generalizability.

\textbf{Statistical analysis}

Mean and standard deviation were calculated for continuous variables. For categorical variables, frequency and percentage was calculated. As more than 80\% of the patients in our data have zero adverse events. We fit a zero-inflated model to consider the number of SAEs as the outcome and the cPTG score as a predictor. We also considered the following variables as controlling covariates in the model: the number of Bevacizumab procedure, days of follow up from latest Bevacizumab procedure, and patient demographics including age, gender, race, and ethnicity. Since the variance of the outcome is much larger than the mean, the data is over-dispersed. So, we preferred the zero-inflated negative binomial model to a zero-inflated Poisson regression model.\textsuperscript{26}

\textbf{Results}

\textbf{Cohort characteristics}

We identified 2,531 unique patients who had been diagnosed with colorectal cancer and treated with Bevacizumab (out of a total of 39,776 colorectal cancer patients). Among these patients, 2,034 (80.4\%) had no SAE while 497 (19.6\%) had at least one SAE. The average age of the patient at her/his last Bevacizumab procedure is 59 years old. Although male patients are slightly more than female patients, there are more female patients with SAEs. Table 2 shows the characteristics of our target population.
Table 2. Demographic characteristics and outcomes of the target population in OneFlorida.

<table>
<thead>
<tr>
<th></th>
<th>Overall (N=2,531)</th>
<th># of SAEs = 0 (N=2,034)</th>
<th># of SAE &gt; 0 (N=497)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (or Mean)</td>
<td>% (or SD)</td>
<td>N (or Mean)</td>
</tr>
<tr>
<td><strong>Age at last PX (years)</strong></td>
<td>59.09</td>
<td>11.4</td>
<td>59.46</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1,206</td>
<td>47.6%</td>
<td>935</td>
</tr>
<tr>
<td>Male</td>
<td>1,325</td>
<td>52.4%</td>
<td>1099</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>1,069</td>
<td>42.2%</td>
<td>868</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>481</td>
<td>19.0%</td>
<td>377</td>
</tr>
<tr>
<td>Hispanic</td>
<td>528</td>
<td>20.9%</td>
<td>407</td>
</tr>
<tr>
<td>Other</td>
<td>17</td>
<td>0.7%</td>
<td>15</td>
</tr>
<tr>
<td>Unknown</td>
<td>436</td>
<td>17.2%</td>
<td>367</td>
</tr>
<tr>
<td><strong>cPTG</strong></td>
<td>0.49</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td><strong>Number of SAEs</strong></td>
<td>0.98</td>
<td>4.08</td>
<td>0</td>
</tr>
<tr>
<td><strong>Follow up days</strong></td>
<td>103.1</td>
<td>66.97</td>
<td>99.95</td>
</tr>
<tr>
<td><strong>Number of Bevacizumab PXs</strong></td>
<td>10.78</td>
<td>11.63</td>
<td>9.96</td>
</tr>
<tr>
<td><strong>First PX to Last PX in days</strong></td>
<td>263.52</td>
<td>328.17</td>
<td>234.75</td>
</tr>
</tbody>
</table>

*PX: Procedure; SAE: Serious Adverse Event; SD: Standard Deviation; cPTG: composite patient-trial generalizability

**Analysis of Bevacizumab colorectal cancer trial eligibility criteria**

From the collection of 57 Bevacizumab colorectal trials, we extracted 1,674 eligibility criteria (i.e., 951 inclusion criteria and 723 exclusion criteria) from ClinicalTrials.gov. 124 of the 951 (13.04%) inclusion criteria contained negations; and 19 of the 723 (2.63%) exclusion criteria contained negations. On average, each colorectal trial has 26 (4 to 65) inclusion criteria and 27 (0 to 39) exclusion criteria. Comparing to our previous study on Hepatitis C Virus (HCV) trials, colorectal cancer trials have significantly more inclusion and exclusion criteria (5.56 inclusion criteria and 7.98 exclusion criteria in HCV trials). We then extracted the core elements of each inclusion/exclusion criterion and summarized these 1,674 eligibility criteria into 678 unique criterion patterns. Many of the inclusion and exclusion criterion patterns were fundamentally similar (i.e., querying the same core data elements). Note that some criteria can be decomposed into multiple sub-criteria; we thus considered the smallest units as individual study traits (e.g., “history of primary CNS (central nerve system) tumor, or stroke” can be decomposed into “history of primary CNS tumor” and “history of stroke”). Table 3 shows the top 10 most frequent criterion patterns, separated by inclusion vs. exclusion.

**Table 3. Top 10 frequent criterion patterns used by the 57 trials, separated by inclusion vs. exclusion.**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Inclusion Criterion Pattern</th>
<th>Study Coverage # of Studies (%)</th>
<th>Exclusion Criterion Pattern</th>
<th>Study Coverage # of Studies (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aspartate aminotransferase (AST)</td>
<td>32 (56.14%)</td>
<td>Unstable angina</td>
<td>36 (63.16%)</td>
</tr>
<tr>
<td>2</td>
<td>Measurable disease</td>
<td>29 (50.88%)</td>
<td>Myocardial infarction</td>
<td>34 (59.65%)</td>
</tr>
<tr>
<td>3</td>
<td>Age</td>
<td>29 (50.88%)</td>
<td>Radiotherapy</td>
<td>30 (52.63%)</td>
</tr>
<tr>
<td>4</td>
<td>Absolute neutrophil count</td>
<td>29 (50.88%)</td>
<td>Congestive heart failure</td>
<td>26 (45.61%)</td>
</tr>
<tr>
<td>5</td>
<td>Platelets</td>
<td>27 (47.36%)</td>
<td>Pregnant</td>
<td>22 (38.60%)</td>
</tr>
<tr>
<td>6</td>
<td>Metastatic colorectal cancer</td>
<td>27 (47.36%)</td>
<td>Bone fracture</td>
<td>22 (38.60%)</td>
</tr>
<tr>
<td>7</td>
<td>Hemoglobin</td>
<td>25 (43.86%)</td>
<td>Significant traumatic injury</td>
<td>22 (38.60%)</td>
</tr>
<tr>
<td>8</td>
<td>Bilirubin</td>
<td>22 (38.60%)</td>
<td>Bleeding diathesis</td>
<td>21 (36.84%)</td>
</tr>
<tr>
<td>9</td>
<td>Creatinine</td>
<td>22 (38.60%)</td>
<td>Chemotherapy</td>
<td>20 (35.09%)</td>
</tr>
<tr>
<td>10</td>
<td>Alanine transaminase (ALT)</td>
<td>20 (35.09%)</td>
<td>Skin ulcers</td>
<td>20 (35.09%)</td>
</tr>
</tbody>
</table>

However, not all eligibility criteria were computable against our OneFlorida patient database. We found that 194 (28.61%) of the 678 unique patterns were not computable. The main reasons are: (1) the criterion asked for subjective information (e.g., patient’s consent or investigator’s judgement of patient’s health status); and (2) the data elements needed for the criterion were not presented in the OneFlorida data (e.g., “performance status ecog 0-1” is not captured).

The composite patient-trial generalizability (cPTG) score
Out of the 57 trials, there are 10 Phase I, 6 Phase I/II, 31 Phase II, and 7 Phase III trials (the other 2 have no phase information). We randomly selected 4 trials out of the 7 Phase III trials to model patient eligibility and trial generalizability. We selected Phase III trials as they are conducted to expand on the safety and effectiveness results from Phase I and II trials, to compare the drug to standard therapies, and to evaluate the overall risks and benefits of the treatment, right before the treatment can be approved by the FDA and be put on the market. A Phase III trial typically recruits larger groups of people with more relaxed eligibility criteria comparing to Phase I and II trials.

To calculate cPTG scores, we first calculated the mGIST scores of the 4 trials using population A as the target population (i.e., patients who were treated with Bevacizumab for colorectal cancer). Table 4 lists the total number of study traits based on the eligibility criteria (regardless of inclusion or exclusion), the number of computable traits, and the mGIST score for each of the 4 trials. There are 219 unique study traits across the 4 trials, and the top 5 common traits are: age, aspartate aminotransferase (AST), bone fracture, pregnancy test, and skin ulcers.

Table 4. Characteristics of study traits, mGIST scores, and the relationships between eligibility and number of SAEs.

<table>
<thead>
<tr>
<th>Trial ID</th>
<th>Total # of traits</th>
<th># of computable traits</th>
<th>mGIST</th>
<th>Mean # of SAEs</th>
<th>Wilcoxon Rank Sums</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Eligible (SD)</td>
<td>Not eligible (SD)</td>
</tr>
<tr>
<td>NCT00025337</td>
<td>46</td>
<td>38</td>
<td>0.547</td>
<td>0.8 (3.0)</td>
<td>1.5 (5.8)</td>
</tr>
<tr>
<td>NCT00252564</td>
<td>84</td>
<td>64</td>
<td>0.750</td>
<td>1.0 (4.1)</td>
<td>1.4 (4.5)</td>
</tr>
<tr>
<td>NCT00096278</td>
<td>88</td>
<td>66</td>
<td>0.584</td>
<td>0.9 (4.1)</td>
<td>1.4 (4.4)</td>
</tr>
<tr>
<td>NCT00112918</td>
<td>82</td>
<td>57</td>
<td>0.307</td>
<td>0.5 (2.0)</td>
<td>1.2 (4.6)</td>
</tr>
</tbody>
</table>

*SAE: serious adverse event; SD: standard deviation; mGIST: multi-trait Generalizability Index on Study Traits (GIST 2.0).

We first used Wilcoxon rank sums test to test the difference of the number of SAEs between eligible vs. no eligible patients and found the differences are statically significant for all 4 trials (Table 4).

The relationship between the patient-trial generalizability and clinical outcomes

The zero-inflated negative binomial model results have two parts, as shown in Table 5.

Table 5. Zero-inflated negative binomial model for the relationships between cPTG and clinical outcomes.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Wald 95% Confidence Interval</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part 1: logistic part for excessive zero (i.e., having no SAE)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at last Bevacizumab PX</td>
<td>0.03</td>
<td>(-0.01, 0.06)</td>
<td>0.1153</td>
</tr>
<tr>
<td>Female vs Male</td>
<td>-0.78</td>
<td>(-1.4, -0.10)</td>
<td>0.0242  &lt; 0.05</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic vs NHW</td>
<td>-0.55</td>
<td>(-1.37, 0.27)</td>
<td>0.1882</td>
</tr>
<tr>
<td>NHB vs NHW</td>
<td>-0.35</td>
<td>(-1.12, 0.42)</td>
<td>0.3767</td>
</tr>
<tr>
<td>Other vs NHW</td>
<td>0.12</td>
<td>(-0.49, 4.33)</td>
<td>0.956</td>
</tr>
<tr>
<td>Unknown vs NHW</td>
<td>0.53</td>
<td>(-0.22, 1.28)</td>
<td>0.1638</td>
</tr>
<tr>
<td>Follow up day</td>
<td>0.00</td>
<td>(-0.01, 0.00)</td>
<td>0.1882</td>
</tr>
<tr>
<td>Number of Bevacizumab PXs</td>
<td>-0.05</td>
<td>(-0.10, -0.01)</td>
<td>0.0262  &lt; 0.05</td>
</tr>
<tr>
<td>cPTG score</td>
<td>1.30</td>
<td>(-0.48, 3.08)</td>
<td>0.1519</td>
</tr>
<tr>
<td>Part 2: negative binomial part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at last Bevacizumab PX</td>
<td>-0.010</td>
<td>(-0.028, 0.009)</td>
<td>0.3038</td>
</tr>
<tr>
<td>Female vs Male</td>
<td>-0.225</td>
<td>(-0.586, 0.137)</td>
<td>0.223</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic vs NHW</td>
<td>-0.262</td>
<td>(-0.713, 0.189)</td>
<td>0.2542</td>
</tr>
<tr>
<td>NHB vs NHW</td>
<td>-0.078</td>
<td>(-0.521, 0.364)</td>
<td>0.729</td>
</tr>
<tr>
<td>Other vs NHW</td>
<td>0.068</td>
<td>(-2.834, 2.970)</td>
<td>0.9633</td>
</tr>
<tr>
<td>Unknown vs NHW</td>
<td>0.057</td>
<td>(-0.450, 0.564)</td>
<td>0.8249</td>
</tr>
<tr>
<td>Follow up day</td>
<td>0.003</td>
<td>(-0.001, 0.006)</td>
<td>0.1047</td>
</tr>
<tr>
<td>Number of Bevacizumab PXs</td>
<td>0.024</td>
<td>(0.012, 0.037)</td>
<td>0.0002  &lt; 0.05</td>
</tr>
<tr>
<td>cPTG score</td>
<td>-1.079</td>
<td>(-1.996, -0.162)</td>
<td>0.0211  &lt; 0.05</td>
</tr>
<tr>
<td>Dispersion</td>
<td>5.421</td>
<td>(3.561, 8.253)</td>
<td></td>
</tr>
</tbody>
</table>
The first part is a logistic model, estimating the probability of being an excessive zero (i.e., having no SAE). Two variables—gender and the number of Bevacizumab procedures—are statistically significant at 0.05 level in this part of the model. The odds ratio for female verse male of being an excessive zero is exp(-0.78) = 0.458. This indicates that female has a higher possibility of having SAEs than male while holding other predictors constant. Further, with the number of Bevacizumab procedures increasing by 1, the odds of having no SAE decreases by 0.049 (i.e., 1 - exp(-0.05)=0.049). This indicates that having more Bevacizumab procedures increases the possibility of having SAEs, holding other predictors constant. Although the cPTG score is not statistically significant in the logistic model, the estimate equals to 1.30, meaning the odds of having no SAE increase by exp(1.30*0.1) - 1 = 0.139 for a 0.1 increase in the patient-trial generalizability score, holding other variables constant.

The second part is a negative binomial regression model, estimating the expected number of SAEs. Two variables—the number of Bevacizumab procedures and cPTG score—are statistically significant at 0.05 level. The expected number of SAEs increases by exp(0.024) – 1 = 0.0243 for 1 increases in the number of Bevacizumab procedures, indicating more Bevacizumab procedures would have more SAEs. Further, the expected change in the number of SAEs decreases by 1 – exp(-1.079*0.1) = 0.102 for 0.1 increase in cPTG. This indicates that a higher cPTG would lead to a smaller number of SAEs, while holding other variables constant. The dispersion is 5.421 (>1); thus, using a binomial model is more appropriate than a Poisson model.26

Discussion and conclusion

Our results successfully connect the generalizability to clinical outcomes using RWD. First, we showed that the patients who are eligible for the original trials (that were used to develop the treatment, population B in our case) have better treatment outcomes compared with those who are not eligible (i.e., population C) in real-world clinical practice. This indicates reduced generalizability of the original trials because of the differences between the characteristics of the trial (or trial-eligible) population and ineligible population. Because of these differences, the original trials may have failed to gather sufficient safety and efficacy data on the entire target population, resulting in poorer outcomes when the treatment is applied in real-world clinical settings. Second, we devised a new patient-trial generalizability score considering both patient-level eligibility and trial-level a priori generalizability. A higher cPTG score of a patient indicates that the patient can be eligible for more (original) trials (i.e., the patient’s characteristics are more similar to those in the trial population) and these trails have a higher a priori generalizability.

These results have multiple implications. First, it shows the feasibility of using RWD, especially EHRs, to assess patients’ clinical outcomes (e.g., adverse events) and examine various patient characteristics associated with these outcomes. Being able to capture AEs using data collected through routine the standard of care provides us the opportunity to create informatics surveillance systems for post-market drug safety monitoring. Comparing to the existing AE reporting systems (e.g., the FDA Adverse Event Reporting System, [FAERS]), EHRs is a much richer data source (e.g., lab results and diagnoses of other diseases) that provides a more complete picture of patients’ characteristics. Further, being able to work with data from a large data research network not only gives us a large sample size and sufficient power to make robust conclusions but also makes our approaches adaptable to other similar networks. Because of the use of a common data model (CDM), our study can be readily replicated in other sites of the PCORnet, which covers more than 100 million patients across the United State.28 Moreover, it is also possible to extend our study to other clinical research network initiatives such as the National Center for Advancing Translational Sciences (NCATS)’s Clinical and Translational Science Awards Accrual to Clinical Trials (CTSA ACT) network and the Observational Health Data Sciences and Informatics (OHDSI) consortium. Even though they are using different CDMs, adapting our approach to a new CDM is rather straightforward.

Further, our study builds a body of evidence to support the development of an eligibility criteria design tool for optimizing study generalizability at the study design phase. Such an eligibility criteria design tool is much needed and will benefit stakeholders of the clinical trial communities. To the extent that trial participants share the same characteristics as the majority of patients, trials should be developed to fit patient profiles seen in real-world clinical practice, and thus, facilitate the application of trial results to real-world clinical settings.29 This could be achieved if trials had few restrictions on eligibility, allowing more representative patients to participate. A significant proportion of real world patients are unable to participate in clinical trials due to stringent exclusion criteria, but many still received treatment outside of clinical trials and benefited from therapy.30 A recent special issue on eligibility criteria in the Journal of Clinical Oncology—a leading clinical oncology research journal—calls for “broadening eligibility criteria to make clinical trials more representative.”31 As a response, the National Cancer Institute (NCI) has revised its clinical trial protocols to expand access for previously excluded patients in an effort to ensure that study participants are more reflective of real-world populations.32 Our study provides initial data evidence that 1) a priori
generalizability is qualifiable and 2) combined with patient eligibility, the cPTG score can predict clinical outcomes in real-world patients. This eventually could lead to metrics that rationalize the design (or relaxation) of eligibility criteria.

Last, our study also leads to opportunities to develop a computable eligibility criteria framework for EHR-based cohort identification to facilitate trial recruitment. One can imagine a tool that a trial investigator could not only identify patients who are eligible to the trial of interest based on their EHRs, but also be able to estimate the impact of recruiting a particular patient on the trial’s generalizability and possible clinical outcomes on similar patients.

Our study is not without limitations. The process of analyzing and decomposing existing eligibility criteria into computable eligibility criteria against EHRs is both time- and labor-intensive. Exploring advanced natural language processing tools tailored for analyzing eligibility criteria might be beneficial. Nevertheless, to tease out the subtle ambiguities in free-text eligibility criteria, human judgments are always needed. Moreover, many eligibility criteria cannot be accurately translated into database queries (e.g., “At least 10 days since prior aspirin dose of more than 325 mg/day”). We took a simplistic approach and did not consider these temporal constraints, which may lead to a small number of inaccurate identifications of patient eligibility. Further, some criteria (e.g., lab results) may be matched to multiple observations (e.g., blood pressure) that vary from time to time. We made an assumption that the patient will meet the criterion as long as if any one of the observations fell into the permissible value range.

In sum, our results are significant. Our ultimate goal is to provide an easy-to-use and more efficient computable eligibility criteria construction platform for investigators to identify eligible patients based on their existing EHRs while maximizing the trial’s a priori generalizability.

Acknowledgements
This work was supported in part by NIH grants R21 AG061431 and UL1TR001427, the OneFlorida Cancer Control Alliance (funded by James and Esther King Biomedical Research Program, Florida Department of Health Grant Number 4KB16), the OneFlorida Clinical Research Consortium (CDRN-1501-26692) funded by the Patient Centered Outcomes Research Institute (PCORI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or PCORI.

References


Abstract

Approximately 60 million people worldwide suffer from epileptic seizures. A key challenge in machine learning approaches for epilepsy research is the lack of a data resource of analysis-ready (no additional preprocessing is needed when using the data for developing computational methods) seizure signal datasets with associated tools for seizure data management and visualization. We introduce SeizureBank, a web-based data management and visualization system for epileptic seizures. SeizureBank comes with a built-in seizure data preparation pipeline and web-based interfaces for querying, exporting and visualizing seizure-related signal data. In this pilot study, 224 seizures from 115 patients were extracted from over one terabyte of signal data and deposited in SeizureBank. To demonstrate the value of this approach, we develop a feature-based seizure identification approach and evaluate the performance on a variety of data sources. The results can serve as a cross-dataset evaluation benchmark for future seizure identification studies.

1 Introduction

Epilepsy is one of the most common neurological disorders, and an estimated 60 million people suffer from this condition worldwide. An epileptic seizure (hereafter referred to as seizure) is a brief episode of signs or symptoms due to transient and undesired, excessive and synchronous electrical discharge, involving large numbers of neurons within the brain. When a seizure occurs, it may change patient behavior, movements or feelings, and level of consciousness. It may produce temporary confusion, uncontrollable jerking movements of the arms and legs, inability to speak, and loss of consciousness or awareness. In a worst-case scenario, frequent seizures may predispose a person to sudden unexpected death of epilepsy (SUDEP). The more frequent the seizures, the higher the risk of SUDEP.

The identification and diagnosis of specific types of epileptic seizures require long-duration monitoring of patients’ electrophysiological status including electroencephalogram (EEG) and electrocardiogram (ECG) in Epilepsy Monitoring Units. However, manual processing of long-duration electrophysiological signals after their acquisition can be a tedious, labor-intensive, time-consuming, and error-prone task. Therefore, it is desirable to develop robust and reliable automated techniques for epileptic event identification based on electrophysiological signals, in order to facilitate real-time seizure monitoring. However, most existing approaches for seizure analysis face the following challenges.

Data preparation. Data preparation, including data extraction, cleaning, reorganization, and reformattting, is an important step in seizure-related studies and applications. In general, an estimated 80% of the work conducted by data scientists is focused on data preprocessing. Cleaning and organizing data take 60% of their time. It is time-consuming and labor-intensive to create a clean, high-quality, and large dataset.

During the data preparation step for the Center for SUDEP Research (CSR), we have encountered several challenges. (1) Data corruption: Sometimes the clinical data may be corrupted during the transaction or conversion process. Certain Zip archives may not be unzipped using the existing software or tools. (2) Inefficiently use of computing resources: A plenty of time and memory were expended on retrieving and extracting irrelevant data from the dataset. (3) Lack of standardized annotation labels: In the CSR data repository, there are three types of labels in an annotation file: official annotation label (OAL), machine label, and free-text label. OALs are from the CSR official and curated annotation label library. Machine labels are created by different machine learning algorithms. Free-text labels are generated by different annotators or experts, which may contain typical manual errors, such as: (i) spelling mistake, e.g., “clinical eventend” should be “clinical event end”; (ii) many-to-one correspondence, i.e., different labels represent the identical meaning for an annotation, e.g., both “clinical end” and “clinical seizure end” represent the end time of a
seizure; and (iii) annotation incompleteness, e.g., some annotations only have the label of the event start time without the label of the event end time.

**Lack of integrated data management, visualization, and analytic system.** The need for data integration and management systems arises with the increasing volume of data. Researchers need query and data exploration tools to find and extract subset of signal data meeting their specific requirements from a large dataset. The lack of an effective and efficient solution to capture, integrate, manage and query seizure-related data becomes one of the main challenges in epilepsy research. Beyond epilepsy research, supporting high-quality data management, analytics, and visualization is one of the goals in National Institutes of Health (NIH) strategic plan for data science.

**Lack of a cross-dataset evaluation benchmark for seizure analysis.** Most previous approaches for seizure analysis were performed and evaluated on a single public seizure dataset, as there is a lack of cross-dataset evaluation benchmark for testing the performance of an approach applied on different datasets. Such cross-dataset evaluation benchmark may empower the development of reliable and robust analytical methods for real-world applications.

In this paper, we introduce SeizureBank, a web-based signal data management and visualization system for epileptic seizure research, to address the above-mentioned challenges. SeizureBank has several key features as follows.

**A scalable seizure data preparation pipeline.** We develop a data preparation pipeline with a novel data model. We address the issues of data corruption and lack of standardized annotation labels with three steps in the pipeline: data extraction, data validation, and data correction. We improve computational performance for data preparation with scalable algorithms, making it more suitable for rapid changes in the growth of data. The input of this pipeline is the raw data from the CSR data repository, including the signal data in European Data Format (EDF) and associated annotation files. The output is the processed data stored in a database with our new data model (Section 3.2.B).

**An analysis-ready electrophysiological signal dataset.** We extract seizure-related signal data according to annotation files in the CSR using our data preparation pipeline, and then import the analysis-ready data into the database. It provides a set of epileptic seizure-related data. The signal fragments (i.e., signal data during a time period) use less storage space in a simple and common file format, such as JavaScript Object Notation (JSON) and Comma Separated Values (CSV). They can be individually or group exported instead of using entire large signal file with complex data structures such as EDF. With such analysis-ready dataset, researchers may obtain the preprocessed data directly and spend less time on data preparation and cleaning.

**A feature-rich, web-based seizure data management and visualization system.** The main idea of this system is to provide a set of functions that enable users to perform a clinical event-specific query and exploration at precise time-points. With such capabilities, researchers can compose complex queries over a number of patients and signal files according to their research requirements, quickly visualize physiological signal fragments of interest, assess the feasibility of research studies, and make appropriate full data access requests for further analysis.

To validate the utility of SeizureBank, we implemented our new data preparation pipeline on a subset of electrophysiological signal data (1.15 TB) from the CSR repository and performed a preliminary performance evaluation to demonstrate the feasibility and advantages of our data preparation process. Based on the analysis-ready seizure signal dataset, we developed a feature-based seizure identification technique. We extracted 134 new features and applied four classifiers to three different datasets: a dataset in SeizureBank and two widely used public datasets. Our experiment results on two public datasets compared favorably to previous work. All the results of the three datasets can serve as a cross-dataset benchmark for future seizure identification research and applications.

**2 Background**

**2.1 The Center for SUDEP Research (CSR)**

CSR is a Center Without Walls Initiative of the National Institute for Neurological Disorders and Stroke (NINDS) for collaborative research in the epilepsies. CSR is comprised of researchers from 14 institutions, bringing together diverse expertise for SUDEP research. CSR provides a comprehensive, curated prospectively constructed repository of epilepsy-related data consisting of electrophysiological signals linked to risk factors and outcome data for about 3,000 epilepsy patients (with a broad spectrum of age, social, racial, and ethnic) with thousands of 24-hours recordings.
2.2 Public epileptic seizure datasets

Two popular seizure datasets are publicly available. The first one is the University of Bonn seizure dataset (UBSD), which is considered a benchmark dataset of testing classification techniques. This dataset contains five subsets (A, B, C, D, and E) of EEG data from both normal and epileptic subjects. Set A and B consist of normal EEG data during an awake state with eyes open and eyes closed respectively from 5 subjects. Sets C, D, and E contain recorded signal data using implanted intracranial electrodes from 5 epileptic patients. However, set C and D contain only activities measured during seizure-free intervals, and only set E carries seizure activities. Each subset contains 100 text files of segmented EEG signals fragments (in 23.6 seconds) and a total of 4,097 samples at a 173.61 Hz sampling rate. There is no available information about the age and gender associated with this dataset.

The second public dataset is the CHB-MIT scalp EEG database. EEG records at a sampling rate of 256 Hz were obtained from 23 pediatric patients (5 male patients in the age group of 3 to 22 years, and 18 female patients in the age group of 1.5 to 19 years). The basic statistics of the two datasets are summarized in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>UBSD</th>
<th>CHB-MIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>NA</td>
<td>1.5-22 years</td>
</tr>
<tr>
<td>Gender</td>
<td>NA</td>
<td>Male (5), Female(18)</td>
</tr>
<tr>
<td>Number of Patients</td>
<td>10</td>
<td>23</td>
</tr>
<tr>
<td>Total Seizure Time</td>
<td>2,360s</td>
<td>11,098s</td>
</tr>
<tr>
<td>Total Non-seizure Time</td>
<td>9,440s</td>
<td>2,849,107s</td>
</tr>
</tbody>
</table>

3 Methods

We introduce a seizure analysis workflow with SeizureBank, including a new data preparation pipeline and a data management and visualization system (Figure 1). With this workflow, we create an analysis-ready seizure-related signal dataset and apply different seizure identification approaches to a variety of data resources. First, we obtain raw data from the CSR data repository (Figure 1.A). Then, with our data preparation approach (Figure 1.B, details are described in Section 3.1), the analysis-ready and well-formatted (in CSV or JSON) seizure-related electrophysiological signal data are imported into the database. With a web-based system dedicated for SeizureBank, researchers can query, visualize, and download each seizure signal fragment of interest with built-in signal visualization and analytic functions (Figure 1.C, as described in Section 3.2). With our new signal dataset, we further introduce and evaluate different seizure identification approaches in Section 3.3 to demonstrate a pathway for constructing a cross-dataset evaluation benchmark based on three datasets, including the UBSD, CHB-MIT, and SeizureBank, for future research.

3.1 Data preparation pipeline in SeizureBank

A. Data extraction. In the CSR data repository, about 10% of raw data had data corruption or unzipping issue. It is a serious problem during the data preparation since data parsing may be interrupted due to "uncompress file error" or "unknown file exception." It is difficult to manually fix the corrupted data. We develop three functions to recover uncorrupted files from a corrupted zip archive. There is no need to read and uncompress the entire archive, especially for signal files, to obtain the required information. We extract the metadata and annotation data, which is only a small part (< 0.1%) of the entire file, from the archive. Based on the extracted metadata and annotation data, we filter the irrelevant files and focus on the signal files with seizure annotations. This way, the memory usage and the running time for data extraction can be improved. It becomes more efficient to retrieve and extract a piece of the signal containing useful data. For label matching, we use Levenshtein distance to handle the misspelling of labels. Informally, the Levenshtein distance between two words is defined as the minimum required number of single-character edits, such as insertions, deletions or substitutions, to change one word into another one. Thus, our program can recognize the majority of labels with few exceptions. We create two mapping files, named OAL mapping and OAL alias mapping, for label pairing and pattern matching to address many-to-one label issue. Using the two mapping files, the program
can find the correct corresponding labels based on the mapping files.

B. Data validation. All the extracted seizure-related label-pairs from step A, containing the start time and end time of a seizure event, are stored in an annotation database. It is necessary to verify the correctness of extracted label-pairs by domain experts due to the issue of inconsistent annotation labels. There are three potential cases as follows. (1) In the case of correct label-pairs, we extract and reformat the corresponding signal data to an efficient data format and import the reformatted data into the database (Figure 1.B.4). (2) In the case of incomplete label-pairs caused by missing labels, we record all these incomplete label-pairs, which can not find the corresponding start time label or end time label, and related annotation files. Experts manually verify the annotation labels with the signal and annotation files and recover the missing labels. Then we add recovered annotation labels into the mapping file to re-extract label-pairs (Figure 1.B.5). (3) In the case of anomalous label-pairs caused by annotation incompleteness issue, we correct seizure data in the step of data correction (Figure 1.B.6) described next.

C. Data correction. There may exist two types of anomalous seizure label-pairs. (1) Caused by incorrect start or end timestamp. For example, if the duration of a label-pair is more than 60 minutes, which rarely occurs in an epileptic seizure, the label-pair will be recorded and sent to domain experts for further verification. After verification, we correct the timestamp of the label-pair based on experts’ feedback. (2) Caused by overlap. For extracted label-pairs, a label-pair with long duration may contain several label-pairs with short durations. We divide the long one into several separate short-duration label-pairs. Overlapping portions between extracted label-pairs are removed afterwards.

The signal fragments are extracted based on curated label-pairs and reformatted with new data format (Figure 1.B.7). All the signal fragments are imported into the database for further querying, visualization, and analysis (Figure 1.B.8).

3.2 Data management and visualization system

SeizureBank provides an integrated data management and visualization system (Figure 1.C) with feature-rich web interfaces for clinicians and researchers, including (1) cloud storage (the storage in the server) for seizure-related data, to address the storage limitations of the local machine; thus, researchers no longer need to store the entire signal dataset and only keep the relevant seizure data for further analysis; (2) standardized, metadata-based search and query of seizure-related signal data for accessing precise signal fragments of a specific patient; (3) interfaces for querying, seizure-specific extraction, and repackaging of electrophysiological signal data; (4) online visualization of single, multiple, or cross-sectional seizure waveforms; (5) signal analytic functions; and (6) customized seizure-fragments exporting function.

A. System architecture. The system architecture of SeizureBank (Figure 2) consists of (1) a back-end database to store signal fragments with its metadata, (2) a query engine based on the Rails framework for processing user requests, such as event-based searching and exporting and signal fragments, and (3) user interfaces and information display. Overall, SeizureBank system uses the Model View Controller (MVC) architecture pattern. It is implemented using Ruby on Rails and Data-Driven Documents (D3) which is an open source JavaScript library for data visualization. We describe different components of SeizureBank in following subsections.

B. Data model. We select MongoDB as the backend database in the server (serving as the cloud) for data storage and management. MongoDB is a NoSQL database system based on JSON-like documents with schemas. MongoDB provides more flexible data modeling, considerable scalability (for large dataset), and better data management performance (high queries per second (QPS) for both reading and writing) compared with SQL databases.

We design and develop three different data models to store patient metadata, signal metadata, and signal fragments. As shown in Figure 2, patient metadata model consists of the information of each patient, such as age and gender. Signal metadata model is designed for storing metadata of each signal fragment, such as the sampling frequency and
seizure duration. Signal fragment model contains the raw physiologic signal data of each clinical event, for example, the epileptic seizure event. The main idea of our new data schema is to break large sequential data files into minimal fragments. Such fragments can be assembled, retrieved, rendered, or repackaged according to users’ requests. Based on these models, seizure-related data can be searched, viewed, and exported in various ways for different purposes.

C. Interface. The interface of SeizureBank incorporates a set of query and visualization functionalities.

Query and exporting interface. SeizureBank query interface is designed as a feature-tag based item search engine. It provides several attributes and features for the user to choose, such as age range, gender, different types of signal channels, events types, and seizure duration. These choice options are implemented as several simple drop-down lists for users to construct queries. Based on user-selected features, the query results include all the metadata of corresponding patients and are shown in the results table. Furthermore, users can view detailed information about each patient. For data exporting, based on the query, users can customize specific clinical and signal data in the downloaded file using CSV format instead of downloading entire large signal files in EDF.

Visualization interface. SeizureBank provides a new interactive visualization interface based on a similar idea in our previous work17. It not only supports all the features in the previous work, such as fragment navigator, functional time-scale scrollbar, signal color configuration, and amplitude-scale scrollbar, but also integrates several new signal visualization and analytic functions, such as graphical montage configuration, signal Fourier analysis and spectrogram analysis18.

3.3 Feature-based seizure identification/classification approach

Leveraging the analysis-ready seizure-related signal dataset in SeizureBank, we develop a feature-based seizure identification technique for epileptic seizure research and evaluate the performance on a variety of datasets.

A. Feature extraction. We extract 134 features including time-domain, frequency-domain, non-linear, and wavelet-based features: (1) five time-domain features19,20 consisting of kurtosis, skewness, and Hjorth parameters; (2) 24 frequency-domain features including spectral power and spectral amplitude of 12 frequency bands (slow-oscillations: 0.5-2.0Hz, slow-wave activity: 0.5-5.5Hz, delta: 1-4Hz, theta: 4-8Hz, alpha: 8-10.5Hz, spindle: 10.5-14.5Hz, sigma: 12-15Hz, slow sigma: 12-13.5Hz, fast sigma: 13.5-15Hz, beta1: 15-20Hz, beta2: 20-30Hz, gamma: 30-60Hz); (3) three non-linear dynamics features were extracted, namely, C0 Complexity21, Shannon Entropy22, and Spectral Entropy23; and (4) 102 wavelet-based features: wavelets are a comparatively recent approach for signal processing, and the main advantage is that wavelets allow multi-resolution analysis in time and frequency, simultaneously24. EEG signals are subjected to 5-level decomposition using Daubechies-4 wavelet. From the decomposition process, a total of 6 coefficient sets were generated, and we calculated 17 measurements (Table 2) as wavelet-based features for each coefficient set.

B. Classifier description. Four different classifiers are applied to the extracted features: K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Support Vector Machine (SVM), and Random Forest (RF). KNN is an instance-based learning approach for classifying objects based on the closest training examples, and it is among the simplest of all machine learning models25. ANN contains a set of interconnected processing nodes, called artificial neurons, and connections between artificial neurons; such neurons and connections have weights to adjust learning process and work in unison to solve specific problems25,26. SVM is a supervised learning model and performs classification by maximizing the margin among the classes25. It has been used in various classification problems and regarded as the benchmark classifier in the performance evaluation. RF is an ensemble learning method for classification by constructing a group of decision trees, and its overall output is determined by selecting the classification with the most weighted votes26. The detailed comparison of these classifiers can be found in our previous work26.

C. Evaluation measurements. We evaluate the seizure identification performance of the four classifiers using the

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>the mean of coefficient</td>
</tr>
<tr>
<td>Median</td>
<td>the median of coefficient</td>
</tr>
<tr>
<td>Std</td>
<td>the standard deviation of coefficient</td>
</tr>
<tr>
<td>Max</td>
<td>the maximum of coefficient</td>
</tr>
<tr>
<td>Min</td>
<td>the minimum of coefficient</td>
</tr>
<tr>
<td>Range</td>
<td>the value range of coefficient</td>
</tr>
<tr>
<td>Mad-1</td>
<td>the mean absolute deviation of coefficient</td>
</tr>
<tr>
<td>Mad-2</td>
<td>the median absolute deviation of coefficient</td>
</tr>
<tr>
<td>Norm-2</td>
<td>the Euclidean norm of coefficient</td>
</tr>
<tr>
<td>Skewness</td>
<td>the skewness of coefficient</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>the kurtosis of coefficient</td>
</tr>
<tr>
<td>C0 Complexity</td>
<td>the C0 Complexity of coefficient</td>
</tr>
<tr>
<td>Shannon Entropy</td>
<td>the Shannon Entropy of coefficient</td>
</tr>
<tr>
<td>Spectral Entropy</td>
<td>the Spectral Entropy of coefficient</td>
</tr>
<tr>
<td>Energy</td>
<td>the squared sum of coefficient</td>
</tr>
<tr>
<td>Norm-Engery</td>
<td>normalized energy</td>
</tr>
<tr>
<td>Norm-Std</td>
<td>the normalized Std</td>
</tr>
</tbody>
</table>
recall ($R$), precision ($P$), F1-score ($F1$), and accuracy ($ACC$)$^{27}$. In addition, Cohen’s kappa coefficient$^{28}$ is an important and more robust measure compared with simple percent agreement calculation to evaluate the performance of a classifier. The interpretation of kappa coefficient has six levels of agreement: poor (less than 0), slight (0.00-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.6-0.80), and excellent (>0.8)$^{28}$.

4 Results

SeizureBank was used to process electrophysiological data collected from consented subjects enrolled in the University Hospitals Cleveland Medical Center (UH) dataset in the CSR. In the following, we introduce the dataset in SeizureBank; present the interface results of SeizureBank system, including query, exporting and visualization interface; and then describe the comparison of seizure identification performance of each approach tested on different datasets.

4.1 Data preparation and dataset description

Using our data preparation pipeline, we extracted 224 seizure fragments of 115 patients (41 males, 68 females, and 6 patients with unknown gender) in the age group of 2 to 75 years. A new dataset containing about 28.02 GB of clean and well-structured seizure-related data were extracted from 1.15 TB of EDF files in the UH dataset. The new dataset includes two different types of data: seizure signal fragments, and 2) non-seizure signal fragments which is randomly selected from patients’ recordings. All the detailed characteristics of the new dataset are described in Table 3. Figure 3 shows a summary of basic demographic and clinical data of the patients in SeizureBank: (a) age distribution, (b) gender distribution, (c) distribution of the number of seizures per patient, and (d) distribution of seizure duration. We compared the performance of our previous data preprocessing method and new data preparation pipeline in Figure 4. The execution time of data preparation with new strategy is significantly improved, and the new pipeline shows higher scalability than the previous method when increasing the dataset size. For memory usage, the previous method extracted all signal files in an archive and cost 30% more memory compared with the new method. For other limitations, the previous method can not address the issues of data corruption and annotation label inconsistency.

4.2 System interfaces

A. Query and exporting interface. The features of the query and exporting interface are as follows. (1) Clinical feature selection: users can select appropriate clinical features, such as age and gender, to obtain the data of expected patients; (2) Signal feature selection: users can retrieve specific signal recordings according to selected signal features, such as event type, signal channel, and fragment durations, to better navigate through the data; and (3) Customized signal fragments downloading function. Figure 5 shows the components in query and exporting interface, including (A) feature selection, (B) feature management, (C) query results table, and (D) subject details table. In the feature selection component, users can select specific clinical and signal features. Selected features are editable in feature management. The query are executed based on selected features, and the results are shown in query results table. Clicking the “show” button (marked with a circle) of each patient will display the details in the subject details table. These two functional tables provide a simple way for users to navigate through the data, such as table size control, page-turning, and sorting function. Clicking “download” button enables the user to download corresponding files of

<table>
<thead>
<tr>
<th>Table 3: Description of SeizureBank dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Number of Patients</td>
</tr>
<tr>
<td>Age Group</td>
</tr>
<tr>
<td>Number of Channels</td>
</tr>
<tr>
<td>EEG Type</td>
</tr>
<tr>
<td>Sampling Frequency</td>
</tr>
<tr>
<td>Number of Seizures</td>
</tr>
<tr>
<td>Total Time of Seizures</td>
</tr>
<tr>
<td>Number of Non-seizures</td>
</tr>
<tr>
<td>Total Time of Non-seizures</td>
</tr>
</tbody>
</table>

Figure 3: Summary of SeizureBank dataset

Figure 4: The performance of previous method and new pipeline with different size of dataset.
selected patients.

**Figure 5:** The query and exporting interface of SeizureBank system.

**B. Visualization interface.** The components of the visualization interface are shown in Figure 6. Signal data with selected channels were visualized in the *interactive visualization interface*. Some of the basic functions, such as functional time-scale scrollbar and amplitude-scale scrollbar, were described in details in our previous work\(^{17}\). The system provides a new function of graphical channel/montage configuration and management in the *channel selection* component. Montage is the voltage difference between two different channels, and the *channel selection* component provides the functionalities for editing customized channels/montages. For frequency analysis, the system provides two approaches: Fourier analysis (as shown in Figure 6.C.1) and time-frequency analysis (i.e. spectrogram, as shown in Figure 6.C.2). To measure the length and amplitude of the interesting part of the signal, we designed and implemented a ruler function, as shown in Figure 6.D. The ruler can be triggered by clicking on the signal rendering panel and following the mouse pointer as users moving it around the signal; the X-axis means the signal length with the unit of seconds and the Y-axis means the signal amplitude with the unit associated with each channel.

**C. Usability evaluation.** In order to evaluate the usability of the user interface, we performed a simple questionnaire-based usability test according to existing criteria\(^{29,30}\) to measure users’ satisfaction with the system. We invited four non-technical users to complete several tasks, such as query by clinical features, query by signal features, and visualize
signal fragments with different montage settings. Then the users answered 7 questions in the questionnaire with a rating score. The rating is on a ten-point scale from “Strongly Disagree” (1) to “Strongly Agree”(10). The questions and evaluation results are shown in Table 4. The preliminary results indicate that SeizureBank provides an easy-to-use query and visualization tool, and it is straightforward for users to learn how to use the functionalities in SeizureBank.

4.3 Comparative evaluation of seizure identification approaches

For each of the three datasets (UBSD, CHB-MIT, and SeizureBank), we processed, filtered, and extracted 134 features in the time domain, frequency domain, wavelet fields, and non-linear fields of single EEG C3 channel. We randomly separated all seizure and non-seizure data into two parts: training set (70%) and testing set (30%), and for both sets, the number of seizure fragments was equal to the number of non-seizure fragments. We performed the whole identification procedure 10 times. The average identification evaluation results of three different datasets are shown in Table 5. All the classifiers were implemented with built-in functions in Matlab 2018a, and the parameter settings of each classifier are as follows. (1) For KNN, number of neighbors was 20; (2) For SVM, logical scalar setting was “true” and other settings used default values; (3) For RF, the number of trees was 500 and other settings used default values; and (4) For ANN, the number of hidden layer neurons was 500 and other settings used default values.

The highest value of each measurement among different classifiers is highlighted. The UBSD provided three categories of data: healthy (set A and B), seizure-free (set C and D), and seizure (set E) data. Therefore, there are four distinct data combinations: [seizure, seizure-free], [seizure, healthy], [seizure, seizure-free, healthy], and [seizure, the combination of seizure-free and healthy]. In total, there were 9 test cases in UBSD: (1) A and E; (2) B and E; (3) C and E; (4) D and E; (5) A, C, and E; (6) A, D, and E; (7) B, C, and E; (8) B, D, and E; and (9) A, B, C, D as non-seizure and E as seizure. For all the tests, both SVM and RF had similar excellent performance. RF provided the best performance on the classification of seizure, seizure-free, and healthy data (in the most of tests), and all the kappa coefficients were in the excellent agreement level (>0.8). For the CHB-MIT and SeizureBank, there were two types of signal data: seizure and non-seizure data. Similarly, RF provided the best performance on both CHB-MIT (with 86.20% F1-score and 0.7357 Kappa) and SeizureBank (with 87.35% F1-score and 0.7547 Kappa) dataset. The evaluation results of our approach tested on the UBSD and CHB-MIT datasets were compared favorably to previous approaches; therefore, the identification results applied to SeizureBank could be considered as a benchmark for further seizure identification/classification research using SeizureBank dataset.

All the experiments were executed on a machine with Intel Core i7 2.93 GHz processor, 16 GB main memory, and operating system of macOS Sierra 10.12.6.

5 Discussion

With the growing trend of multi-center collaborative studies, there is a need to adopt emerging data repositories with integrated system to meet research requirements. The results of our data preparation pipeline and system interface demonstrate advantages of SeizureBank for obtaining signal data for epileptic seizure research. Features of SeizureBank can lead to reduced time (less data preparation execution time), space (lower storage requirements), and labor costs (easy to search and query) for analysis-ready seizure-related signal data. The majority of data scientists regard cleaning and organizing data as the least enjoyable work. With SeizureBank, researchers may obtain analysis-ready data directly and spend less time on data preparation and cleaning. SeizureBank is already used in several seizure-related studies in our group, including seizure detection and seizure prediction. The framework including data model, data preparation pipeline, and system architecture of SeizureBank can be extended to other medical and clinical studies using electrophysiological signal data. With the dataset in SeizureBank, we introduced a cross-dataset benchmark for
seizure identification, which can serve as a reference for developing and testing robust and reliable epileptic seizure identification approaches and applications in the future.

Limitations. In this paper, we only extracted seizure-related data from a part of one dataset in the CSR repository. In future work, we will expand more data from additional datasets in the CSR repository. To increase the data diversity of SeizureBank, we plan to extract and import seizure subtype data into our data repository. For data preparation, reading a large signal file (e.g., more than 20 GB) may require more RAM, and we plan to develop a parallel approach for reading and processing large-scale EDF files. For data visualization, we plan to design and implement additional functionalities in our system such as the functionality of user-contributed annotation visualization, where users can store and visualize their identification results.

6 Conclusion

Electrophysiological signal data, such as EEG and ECG, are often used in monitoring, identification, prediction, and therapy of the epileptic seizures. However, a key challenge in epilepsy research is the lack of analysis-ready seizure-specific signal data of large patient cohort repository as well as a system for managing, visualizing, and exporting seizure events. To address this challenge, we developed SeizureBank, an analysis-ready data repository along with web-based data management and visualization interfaces for epileptic seizure research. Our preliminary evaluation demonstrated the feasibility of SeizureBank. With SeizureBank, we also introduced a feature-based seizure identification approach that was tested on multiple datasets. The evaluation results compared favorably to previous seizure identification approaches. Results obtained on the UBSD, CHB-MIT, and SeizureBank datasets can be considered as a cross-dataset evaluation benchmark of seizure identification for future research.

Acknowledgment. This work is supported in part by the NINDS-funded Center Without Walls for Collaborative Research in the Epilepsies U01 awards (U01NS090408 and U01NS090405), University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR001998), and National Science Foundation under MRI Award No.1626364.

References


Predicting Wait Times in Pediatric Ophthalmology Outpatient Clinic Using Machine Learning

Wei-Chun Lin, MD, MS\(^1\), Isaac H. Goldstein, BA\(^2\), Michelle R. Hribar, PhD\(^1\), David S Sanders, MD\(^{2,3}\), Michael F. Chiang, MD\(^{1,2}\)

Departments of \(^1\)Medical Informatics and Clinical Epidemiology and \(^2\)Ophthalmology, OHSU, \(^3\)Legacy Devers Eye Institute, Portland, OR

Abstract

Patient perceptions of wait time during outpatient office visits can affect patient satisfaction. Providing accurate information about wait times could improve patients’ satisfaction by reducing uncertainty. However, these are rarely known about efficient ways to predict wait time in the clinic. Supervised machine learning algorithms is a powerful tool for predictive modeling with large and complicated data sets. In this study, we tested machine learning models to predict wait times based on secondary EHR data in pediatric ophthalmology outpatient clinic. We compared several machine-learning algorithms, including random forest, elastic net, gradient boosting machine, support vector machine, and multiple linear regressions to find the most accurate model for prediction. The importance of the predictors was also identified via machine learning models. In the future, these models have the potential to combine with real-time EHR data to provide real time accurate estimates of patient wait time outpatient clinics.

Introduction

In recent years, hospital systems have increasingly emphasized quality of care, which includes patient satisfaction.\(^{1,2}\) Patients’ perception of wait time in a primary care or specialty care outpatient clinic contributes to patient satisfaction.\(^{3-5}\) In fact, the literature suggests that the time a patient spends waiting for their scheduled appointment is the largest source of patient dissatisfaction.\(^5\) Longer wait time negatively impacts the patient’s satisfaction and distorts the patient’s perception of the quality of care and the physician’s abilities.\(^1\) Although waiting may be unavoidable in outpatient clinics, it is important to note that providing accurate information about wait times could improve patients’ satisfaction by reducing uncertainty.\(^4,7\) However, there are limited studies of proposed models to predict outpatient clinic wait time by using statistical methods.

The complexity of clinic workflows can make predicting patients’ wait time challenging. For example, in an ophthalmology clinic, ophthalmologists typically utilize multiple exam rooms simultaneously, examine patients at different stages in the visit, and integrate ancillary staffs and trainees into the clinical workflow.\(^8\) Thus, it is difficult to provide accurate estimates of wait times in clinical settings. To bridge this gap, secondary use of electronic health records (EHRs) with machine learning algorithms can be a reasonable choice. Supervised machine learning algorithms are an effective tool for predictive modeling with large and sophisticated data sets. These algorithms resist abnormal outliers, rank the relative importance of variables, and self-modify without human supervision.

Our study was performed in an academic ophthalmology department because it is a high volume, fast-paced specialty where estimating patient wait time is paramount. The purpose of this work was to develop analytical models that provide an accurate prediction of patients’ wait time in ophthalmology outpatient clinics using advanced machine learning methods. In addition, we wanted to determine the features that are most important for these predictive models.

Methods

This study was approved by the Institutional Review Board at Oregon Health and Science University (OHSU). OHSU is a large academic medical center in Portland, Oregon. This study was conducted at Casey Eye Institute, OHSU’s ophthalmology department serving all major ophthalmology subspecialties. The department performs over 130,000 outpatient examinations annually and is a major referral center in the Pacific Northwest and nationally. In 2006, OHSU implemented an institution-wide EHR (EpicCare; Epic Systems, Verona, WI) to handle all ambulatory practice management, clinical documentation, order entry, medication prescribing, and billing.
Data Preparation
We used 6 years (January 1, 2012 to March 31, 2018) of office visit data from seven pediatric ophthalmology faculty providers at OHSU Casey Eye Institute. Time-stamp and related data from office visits were extracted from the enterprise-wide clinical warehouse and audit log timestamp data were used to calculate time-related variables. The appointment length is determined by the difference between the office visit check-in and check-out times recorded in the EHR. The provider-patient interaction times were determined by audit log values using previously validated methods. The wait time was defined as total wait time during the clinic appointment (total appointment length – provider-patient interaction time).

Data Preprocessing

Data cleaning – we removed all encounters with missing data and excluded abnormal time variables. Clinic visits were excluded if (1) wait time was longer than 180 minutes, less than 0 minutes, or missing; (2) provider-patient interaction time was longer than 120 minutes or less than 1 minute; (3) appointment length was longer than 300 minutes or less than 10 minutes; (4) arrival interval was longer than 240 minutes or less than 240 minutes.

Data transformation and dimensionality reduction – several variables were re-categorized and all categorical variables were converted to dummy variables. Some potential predictors were removed on the basis of our experience and observation.

Outcome Variables and Analytical Models
The primary outcome variable of regression models was patient wait time, which is a continuous variable. Four machine learning algorithms – random forest, elastic net, gradient boosting machine (GBM), and support vector machine (SVM) and multiple linear regression were developed to predict patient wait time. We chose these models to represent a broad approach to machine learning. Random forest and gradient boosting machine are ensemble decision trees algorithms, elastic net is a regularized method of linear regression, and support vector machine constructs hyperplanes to classify the data.

In addition, patient wait time was transformed into a categorical variable value as “Long” or “Normal”. “Long” indicated the upper half (> 58 minutes) wait time and “normal” indicated bottom half (≤ 58 minutes) wait time. We used the median of the patient wait time as cut-point. Four classification models, including random forest, elastic net with logistic regression, GBM, and SVM, were developed to determine “long wait time” or “normal wait time”. The categorical variable can be used in scheduling templates.

Random forest is an ensemble decision trees algorithm, which grows with a bootstrap sample from the training data. Randomly selected subsets of predictor variables are implemented into each decision tree. The prediction for new observation is made by averaging the output of the ensemble of trees. We used R package randomForest to perform the prediction model.

Elastic net is a regularized regression method that combines penalties of the lasso and ridge methods. We used R package glmnet to build the prediction model.

Gradient-boosting machines is an ensemble learning method for improving predictive performance. Unlike random forest, which builds an ensemble of deep independent trees, GBM makes an ensemble of weak successive trees with each tree learning and improving on the previous one. We used R package gbm, assuming the Gaussian distribution for minimizing squared-error loss.

Support vector machines are supervised learning models used for both classification and regression analysis. The algorithm outputs optimal hyperplanes, which categorizes new examples. We used R package e1071 with a linear kernel to build the prediction model.
Predictor Variables
Twenty-four predictor variables were selected from the original data set and were grouped as below:

(1) Predictors related to date and time: Year, month, day of the month, day of the week, clinic session (AM or PM), clinic hour, which was the hour of the patient’s check-in time relative to the scheduled half-day clinic session.

(2) Patients’ demographic and clinical features: Age, which was grouped as six categories, visit name, patient financial class, whether the patient was a new patient, whether the patient was scheduled or walked in, and International Classification of Diseases, tenth revision (ICD-10) diagnosis code. We grouped the ICD-10 codes into categories based on diagnosis frequency. The 20 most common ICD-10 diagnoses were selected, and all the remaining diagnoses were categorized as “Other” (21 total categories).

(3) Predictors related to clinical examination: Boolean values representing if the patient exam included a pupil dilation, a visual acuity test, a visual field test, a tonometry test, a refraction test, or a fundoscopic exam.

(4) Predictors related to prior visit length: The wait time of patient’s previous office visit and the provider-patient interaction time of patient’s previous office visit.

(5) Predictor for the arrival interval for the current appointment: The arrival interval is defined as the duration between patient check-in time and the scheduled visit time. An arrival interval can either be positive (late arrival) or negative (early arrival).

(6) Predictors related to number of patients: The total number of patients seen by all providers on the day of the appointment and the clinic volume, which indicates the number of patients in the half-day provider’s clinic in which the appointment was scheduled.

Validation and Evaluation of Performance
The data set was randomly assigned to a training set (75%) and testing set (25%) to avoid over-fitting. K-fold cross-validation was used to validate the models. A larger number of folds makes less bias towards overestimating the truly expected error but may have higher variance. In this study, we used 5 fold cross-validation, which shows the flat slope of the learning curve and the size of each validation partition in 5 fold cross-validation being large enough to provide a fair estimate of the model's performance. Each prediction models was developed with the training data set. Root mean square error (RMSE) and R² were used to determine the predictive accuracy of each regression model. We developed five regression models in this study, including random forest, elastic net, GBM, SVM, and multiple linear regressions. Moreover, receiver operating characteristic (ROC) curves and the area under ROC curve (AUC) were used to evaluate the accuracy of the classification models for predicting “long wait time” or “normal wait time”. There are four classification models in the study, including random forest, elastic net with logistic regression, GBM, and SVM. We also ranked the importance of features based on the increase in mean square error of prediction as a result of a certain variable being permuted (%IncMSE) in the random forest regression model. All data processing and analyses were conducted using R programming language version 3.5.2.
Table 1. Descriptive characteristics

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior visit length (minutes)</td>
<td></td>
</tr>
<tr>
<td>Prior wait time</td>
<td>57.26 (30.92)</td>
</tr>
<tr>
<td>Prior interaction time</td>
<td>23.34 (12.10)</td>
</tr>
<tr>
<td>Current arrival interval</td>
<td></td>
</tr>
<tr>
<td>Arrival interval</td>
<td>-10.22 (20.51)</td>
</tr>
<tr>
<td>Patient age</td>
<td>No. (%)</td>
</tr>
<tr>
<td>Baby (&lt; 1 year)</td>
<td>3,444 (9%)</td>
</tr>
<tr>
<td>Toddler (1 – 3 years)</td>
<td>6,187 (16%)</td>
</tr>
<tr>
<td>Pre-School (3 - 6 years)</td>
<td>9,639 (26%)</td>
</tr>
<tr>
<td>School age (6 – 13 years)</td>
<td>12,193 (32%)</td>
</tr>
<tr>
<td>Adolescent (13 – 18 years)</td>
<td>2,969 (8%)</td>
</tr>
<tr>
<td>Adult (&gt; 18 years old)</td>
<td>33,55 (9%)</td>
</tr>
<tr>
<td>Visit name</td>
<td></td>
</tr>
<tr>
<td>Follow-up visit</td>
<td>23,702 (63%)</td>
</tr>
<tr>
<td>Post-operative visit</td>
<td>1,315 (3%)</td>
</tr>
<tr>
<td>Pre-operative visit</td>
<td>481 (1%)</td>
</tr>
<tr>
<td>New patient</td>
<td>8,733 (23%)</td>
</tr>
<tr>
<td>Other</td>
<td>3,556 (10%)</td>
</tr>
<tr>
<td>Patient financial class</td>
<td></td>
</tr>
<tr>
<td>Self-pay</td>
<td>390 (1%)</td>
</tr>
<tr>
<td>Medicaid</td>
<td>15,249 (40%)</td>
</tr>
<tr>
<td>Medicare</td>
<td>1,223 (3%)</td>
</tr>
<tr>
<td>Commercial insurance</td>
<td>16,417 (44%)</td>
</tr>
<tr>
<td>Other</td>
<td>4,508 (12%)</td>
</tr>
<tr>
<td>Pupil dilation</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>20,161 (53%)</td>
</tr>
<tr>
<td>No</td>
<td>17,626 (47%)</td>
</tr>
</tbody>
</table>

Results

Descriptive Data

Table 1 shows the characteristics of the patient visits including patient’s age, visit name, financial class, whether the patient was a new patient, and information about visit length. There were 37,787 (98.8%) patients visits that met inclusion criteria and 445 office visits were excluded (1.2%). Most patients were school age (32%) and pre-school age (26%). Amblyopia, Esotropia, and Exotropia are the most common diagnoses. Most patients used Medicaid and commercial insurance. Besides, 63% of patients came to follow-up visits, 23% of patients were new patients, and approximately half of patients (53%) were scheduled for pupil dilation.

Table 2. Comparison of $R^2$ and RMSE of five predictive models

<table>
<thead>
<tr>
<th>Method</th>
<th>R-squared</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forest</td>
<td>0.3822</td>
<td>24.24</td>
</tr>
<tr>
<td>Elastic net</td>
<td>0.3691</td>
<td>24.74</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.3646</td>
<td>25.01</td>
</tr>
<tr>
<td>GBM</td>
<td>0.3712</td>
<td>24.54</td>
</tr>
<tr>
<td>SVM</td>
<td>0.3633</td>
<td>24.71</td>
</tr>
</tbody>
</table>

Evaluation Performance of Models

The performance of the five models is presented with $R^2$ and RMSE in Table 2. We found that the random forest model had the highest $R^2$ (0.3822), which means the random forest model can explain about 38 percent of the variability of the patient wait time. Also, the random forest model showed the best RMSE (24.24). Thus, we selected the random forest as the best predictive model in this study. The ten most important features of random forest
regression model and their %IncMSE are shown in Figure 1. We found that patients scheduled for pupil dilation, patients’ primary ophthalmologist, and arrival interval have the strongest impact on the predictive model.

Figure 1: The 10 most important features in the random forest model as determined by %IncMSE.

Figure 2 shows the ROC curves of four machine-learning models in the testing dataset for predicting when patient wait times will be long. Area Under the Curve, (AUC) was measured in all models. The random forest model presented the best performance (AUC=81.55% [95% CI 80.69%-82.41%]) followed by the support vector machine (AUC=81.14% [95% CI 80.28%-81.99%]), gradient boosted machine (AUC=80.05% [95% CI 79.17%-80.92%]) and elastic net (AUC=79.88% [95% CI 79%-80.76%]).

Figure 2: ROC curves of machine learning models
Discussion
In this study, we evaluated the applicability of machine learning models to predict patient wait time in pediatric ophthalmology clinics. The key findings from our study were (1) Machine learning models (such as random forest) can accurately predict patient wait time in pediatric ophthalmology outpatient clinic; (2) Machine learning models can provide insight into the factors associated with patient wait time; (3) Patient wait time predictive model can be a useful tool in managing clinical practices.

In our study, the random forest model provided the most accurate prediction (R² 0.38 and RMSE 24.22), which is higher than the prediction accuracy of the linear model (R² 0.36 and RMSE 25.02). Machine learning algorithms are a great choice for predicting sophisticated and noisy phenomena like patient wait time in a pediatric ophthalmology outpatient clinic. Moreover, random forest and GBM models were able to identify patient factors that are associated with patient wait time. The feature importance of random forest regression model and GBM regression model are similar to each other. The top three important predictor variables identified in both models were pupil dilation, then primary ophthalmologist, and arrival interval. The importance of pupil dilation may be because it takes about 25 minutes for pupils to fully dilate, and patients must wait during this period. In this study, the average wait time for patients with pupil dilation was 68.39 minutes (SD=27.09) and for patients without pupil dilation, the average wait time was 44.54 minutes (SD=30.13). The importance of the physician is because the clinic volume of each physician is different and each physician has their own pace when seeing patients. The range of the average wait time of seven faculty providers was 47.36 minutes to 75.01 minutes. The importance of arrival interval is due to patients arriving at the clinic early, which results in wait times. In our study, patients arrived at the clinic about 10 minutes earlier than their scheduled time on average and the Pearson correlation coefficient between patient wait time and arrival interval is -0.22.

Moreover, in Figure 1, we noticed that the year is the 4th important predictor variable in random forest regression model. The importance of year may be because the clinic policies/procedures changed over time. More investigation is needed to determine why this is important since it probably isn’t relevant for future predictive analyses. Exploring the important predictors can be useful for improving the predictive model in the future; for example, building models for each provider, eliminating dilation time from wait time and adjusting wait time calculations to eliminate bias from early arrivals may improve the accuracy of the wait time predictions. For elastic net, multiple linear regression and SVM regression models, we calculated the regression coefficients, which describes the relationship between each feature and the outcome variable. However, the units vary between the different types of variables and cannot be used as the feature importance directly.

Providing accurate information about wait times could improve patients’ satisfaction by reducing uncertainty. In future, the predictive models could be incorporated with real-time data to provide accurate information about wait times for patients in clinical settings. Even with perfect scheduling, it is difficult to completely eliminate patient wait time, so communicating potential delays to patients is important for maintaining patient satisfaction. In the future, the predictions in this study might be used in real-time to help inform patients of expected wait times while in clinic.

There are several limitations in our study. First, the EHR timestamps do not always accurately capture the provider-patient interaction time, which in turn can produce incorrect wait times. That said, we have previously validated that these provider interaction time timestamps were accurate across a wide range of ophthalmology providers at a single institution. Second, the models might improve with a larger data set and our study was limited at a single institution. Finally, the patterns observed in our study might not be generalizable to other subspecialties within ophthalmology or other healthcare systems. Our intention is to extend and replicate these study methods to different patient and provider populations at different institutions in order to increase the generalizability of our findings.
Conclusion
Patient perceptions of wait times in pediatric ophthalmology outpatient clinics can affect patient satisfaction and quality of care. However, there are not many studies about the efficient ways to predict wait time in the clinic. In this study, we found that supervised machine learning models can provide accurate patient wait time prediction and we were able to identify the factors with the largest contribution to patient wait times. It is important to note that patient satisfaction increases when patients are told about their expected wait time. In the future, we may be able to incorporate real-time scheduling data from the EHR to improve estimates of patient wait-time and scheduling efficiency in a clinic setting.

Acknowledgements
Supported by grants T15LM007088, R00LM012238, and P30EY0105072 from the National Institutes of Health, (Bethesda, MD) and unrestricted departmental support from Research to Prevent Blindness (New York, NY). MFC is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems (Pleasanton, CA), a Consultant for Novartis (Basel, Switzerland), and an equity owner in Inteleretina, LLC (Honolulu, HI).
References

Transfer Learning from BERT to Support Insertion of New Concepts into SNOMED CT

Hao Liu, MS, Yehoshua Perl, PhD, James Geller, PhD
Dept of Computer Science, NJIT, Newark, NJ, USA

Abstract

With advances in Machine Learning (ML), neural network-based methods, such as Convolutional/Recurrent Neural Networks, have been proposed to assist terminology curators in the development and maintenance of terminologies. Bidirectional Encoder Representations from Transformers (BERT), a new language representation model, obtains state-of-the-art results on a wide array of general English NLP tasks. We explore BERT’s applicability to medical terminology-related tasks. Utilizing the “next sentence prediction” capability of BERT, we show that the Fine-tuning strategy of Transfer Learning (TL) from the BERT\textsubscript{BASE} model can address a challenging problem in automatic terminology enrichment – insertion of new concepts. Adding a pre-training strategy enhances the results. We apply our strategies to the two largest hierarchies of SNOMED CT, with one release as training data and the following release as test data. The performance of the combined two proposed TL models achieves an average F1 score of 0.85 and 0.86 for the two hierarchies, respectively.

Introduction

Insertion of new concepts into their proper positions in a terminology is a challenging problem in automatic enrichment of terminologies. Traditionally, for terminologies based on a Description Logic, curators tend to use a reasoner such as Snorocket [1] or HermiT [2] to insert a new concept into a terminology’s hierarchy. Recently, researchers have proposed to use Deep Learning models such as Convolutional/Recurrent Neural Network models [3] to verify an IS-A relationship between a new concept and an existing concept, which will recommend the location of the new concept in the hierarchy. In such an approach, concepts are represented by various language embeddings from Natural Language Processing (NLP) [4-6].

However, training such an ML model from scratch is expensive and time-consuming, as it requires large data sets and extensive computing resources. Instead, it is common and efficient to conduct Transfer Learning from pre-trained language representation models to the task of interest. In this paper, we utilize a new general language representation model called Bidirectional Encoder Representations from Transformers (BERT) [7]. BERT is a pre-trained model which obtains state-of-the-art results on a wide range of NLP tasks. It has been integrated into applications [8, 9] for clinical tasks in biomedical domain. We experiment with two Transfer Learning (TL) strategies, Fine-tuning and Pre-training, from the pre-trained BERT\textsubscript{BASE} model to address the terminology enrichment problem. In Fine-tuning, we train a classifier on top of a BERT\textsubscript{BASE} network with IS-A relationships as training data. This training fine-tunes the weights of the pre-trained BERT\textsubscript{BASE} network to enable it to classify relationships between new concepts and existing concepts. The fine-tuning strategy innovatively utilizes the “next sentence prediction” of BERT, to train BERT to distinguish which pairs of concepts should be connected by IS-A relationships. In Pre-training, we accommodate BERT to medical data, by training it from scratch using terminology data from SNOMED CT.

To measure the performance of the two proposed strategies, we use the two largest hierarchies of the SNOMED CT [10] terminology, the Clinical Finding and the Procedure hierarchies, as our testbed. The SNOMED CT release of July 2017 is used as training data. For testing, we use 911 and 2005 new concepts from the Procedure and Clinical Finding hierarchies of the January 2018 release, respectively. The results of this experiment for the Fine-tuning strategy and for the combined strategy starting with Pre-training and continuing with Fine-tuning are reported.

Background

SNOMED CT is an internationally leading clinical terminology, managed by SNOMED International. The Clinical Finding and the Procedure hierarchy of the January 2018 release of SNOMED CT consist of 111,081 and 57,806 active concepts, respectively. SNOMED CT is released twice every year on January and July. All the content for a given release of SNOMED CT terminology is defined in a “snapshot” file. In addition, a “delta” file identifies the individual changes that occurred between the previous release and the current release. A full history of concepts and relationships that are added, changed, or removed is also provided. By comparing the Procedure and Clinical Finding hierarchies of January 2018 with the previous July 2017 release, we found that 911 new concepts were added into the Procedure hierarchy and 2005 new concepts were added into the Clinical Finding hierarchy and placed in the proper
positions in the hierarchy by SNOMED CT’s curators, based on their structural and semantic definitions stated in Description Logic. Liu et al. [11] proposed a methodology to automatically determine the placement of a new concept in the ontology’s hierarchy if a new concept’s name and one of the concept’s parents is given. The solution is based on training a Convolutional Neural Network (CNN) model to distinguish between those pairs of concepts that are connected by IS-A links and those pairs that are not.

For the language representation model, there are two main research streams: Context-free and Contextual representations. Traditional word embeddings such as word2vec [12], GloVe [13], or fastText [14], are Context-free embeddings, which generate a single "word embedding" representation for each token in the vocabulary. Therefore, they are not likely to capture any word meaning changes caused by surrounding context changes. Contextual models, instead, generate a representation of each word that is based on the other words in the context. Contextual representations can further be categorized into unidirectional or bidirectional.

BERT is the first unsupervised, deeply bidirectional system that outperforms previous methods. BERT is a general-purpose "language understanding" model trained on a large text corpus (like Wikipedia), which can be used for various downstream NLP tasks without heavy task-specific engineering. BERT’s model architecture is a multi-layer bidirectional transformer encoder, based on the original implementation proposed by Vaswani et al. [15]. BERT has advanced the state-of-the-art for several major NLP benchmarks, including named entity recognition on CoNLL-2003 [16], question answering on SQuAD [17], and sentiment analysis on SST-2 [18].

Methods

The main contribution of this paper is suggesting a way to harness the high performance of BERT for a critical task in medical terminology enrichment, in spite of the fact that BERT was not trained with medical literature. One approach is to add medical knowledge to the general knowledge learning of BERT. For this we use the SNOMED CT knowledge, providing a “document” for each concept of SNOMED CT. The second approach is to train BERT to be able to distinguish between concept pairs which should be connected by IS-A relationships and pairs that shouldn’t. This kind of learning utilizes the “next sentence prediction” feature of BERT. In the following we describe the technical issues and the details involved in implementing these ideas.

We experimented with two strategies of using BERT. 1) Fine-tuning BERT\textsubscript{BASE} by supervised training a relationship classifier on top of BERT with concept pairs connected by IS-A relationships and pairs not connected (non-IS-A pairs) taken from SNOMED CT. 2) Pre-training BERT\textsubscript{BASE} with unsupervised concept-based “documents” from SNOMED CT, and then fine-tuning it as an IS-A relationship classifier. We implemented the experiments with Tensorflow [19] and ran the testcases on a computer with two Intel Xeon E5-2630-v4 CPUs with processor speed 2.2 GHz; 128 GB memory per CPU and two Nvidia Tesla P100 “Pascal” video cards with 16 GB RAM per GPU.

BERT\textsubscript{BASE} (12 Transformer layers) and BERT\textsubscript{LARGE} (24 Transformer layers) are two models trained on English Wikipedia (2,500M words) and BooksCorpus [20] (800M words) for one million update steps. Due to limited GPU resources, we only used BERT\textsubscript{BASE} in this experiment, since BERT\textsubscript{LARGE} requires resources currently beyond our high-performance hardware. The configuration parameters of the pre-trained BERT\textsubscript{BASE} model are L=12, H=768, A=12, total Parameters=110M, where L is the number of layers (i.e., Transformer blocks), H is the hidden size, and A is the number of self-attention heads. The feed-forward/filter size is set to 4H, i.e., 3072 for H = 768.

The two strategies of our research are as follows:

Strategy 1

Step 1. Fine-tuning the BERT\textsubscript{BASE} model (Figure 1): We extracted IS-A linked and not linked concept pairs as supervised fine-tuning data from the SNOMED CT July 2017 release. We will refer to these pairs as IS-A and non-IS-A pairs. Then we trained a relationship classifier on top of BERT\textsubscript{BASE} with the IS-A and non-IS-A pairs to obtain the BERT\textsubscript{BASE+CLF} model.

Step 2. Prediction on new release data (illustrated in the rightmost process of Figure 1): We tested the trained BERT\textsubscript{BASE+CLF} model to verify IS-A links and the absence of IS-A links for newly added concepts in the SNOMED CT January 2018 release.

Strategy 2

Step 1. Pre-training the BERT\textsubscript{BASE} model (Figure 2): We preprocessed concept-related information from the July 2017 release to generate documents that were used as unsupervised pre-training data. Then we trained BERT\textsubscript{BASE} with unsupervised concept level data so that the trained BERT\textsubscript{BASE+SNO} model integrated terminology information. Then
we applied the fine-tuning process (of Strategy 1) to train a classifier on top of BERT_BASE+SNO to derive the BERT_BASE+SNO+CLF model.

Step 2. Prediction on new release data (illustrated in the rightmost process of Figure 2): We tested the trained BERT_BASE+SNO+CLF model to verify IS-A links and non-IS-A pairs for newly added concepts in the SNOMED CT 2018 January release.

**Figure 1:** The pipeline for Strategy 1 Fine-tuning. CLF is short for Classifier.

**Figure 2:** The pipeline for Strategy 2 combining Fine-tuning with Pre-training. SNO is short for SNOMED CT.

Details of the specific stages of the two strategies are discussed in the following:

**Fine-tuning strategy**

**Data preparation:** The training samples passed to the Fine-tuning process are a set of IS-A and non-IS-A concept pairs. In SNOMED CT, the IS-A concept pairs are given. Thus, we can use them as a positive sample. On the other hand, the negative sample could be all the non-IS-A pairs of concepts. This creates an imbalance between the positive and negative samples, because there are many more pairs not connected by IS-A links. Thus, we pick non-IS-A pairs for the negative sample as follows. For each IS-A pair (A, B) we look for the siblings C₁, C₂, … Cₖ of B. Then we choose non-IS-A pairs (A, Cᵢ) with i=1, 2, … k. The advantage of such pairs is that they are closely related to the corresponding IS-A pair. This will sharpen the distinction between IS-A and non-IS-A pairs in training. For example, (Crushing injury of back, Crushing Injury) defines an IS-A link, while (Crushing injury of back, Shear injury) is a similar pair that should not be connected by an IS-A link. The reason is that Shear injury is a sibling of Crushing Injury, with the same parent Injury by mechanism.
In the data preparation for Fine-tuning, we first extracted the positive and negative samples, and randomly downsampled the negative sample to the size of the positive sample, at the beginning of each training round. We shuffled our dataset and then used 90% of it for training and kept 10% as the test set. The samples went through three preprocessing steps: Text normalization (E.g., Excision of Reinke’s edema, → excision of reinke’s edema), Punctuation splitting (E.g., excision of reinke’s edema, → excision of reinke’s edema), and WordPiece tokenization (excision of reinke’s edema, → ex #cision of rein #ke’s ed #ema). Then the samples were processed by BERT\textsubscript{BASE}, which performed its own preprocessing, including input embeddings, segment masking, labeling, etc [7]. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings. Due to space limitations, we concentrate on the operations that are essential for Fine-tuning.

For example, Urine xanthine level is the child of Evaluation of urine specimen in the SNOMED CT Procedure hierarchy. The input sequence will be “1 Evaluation of urine specimen (t) Urine xanthine level (t)” This input will be converted as one training instance to “[CLS] evaluation of urine specimen [SEP] urine x ##ant ##hine level [SEP]” as shown in Figure 3(a). The first token of the sequence is the classification embedding ([CLS]), representing a classification label. A special token ([SEP]) is used to separate sentences. Out-of-vocabulary words are split into word pieces and denoted with ##. For example, “xanthine” is denoted as three items “x”, “##ant”, and “##hine.” Similarly, Rubella screening is not a child of Down’s screening – blood test. The input sequence “0 Rubella screening (t) Down’s screening – blood test (t)” will be converted to “[CLS] rub ##ella screening [SEP] down ’ s screening - blood test [SEP]” in Figure 3(b).

**Figure 3:** Fine-tuning data: Preprocessing (a) IS-A and (b) non-IS-A concept pairs

**Fine-tuning the BERT\textsubscript{BASE} model:** We fine-tuned the BERT\textsubscript{BASE} model to predict the IS-A and non-IS-A linking for the concept pairs in the test data. This is similar to a binary sentence-pair classification task. We used the sentence prediction capability of BERT\textsubscript{BASE}, trained as BERT\textsubscript{BASE+CLF}, to predict IS-A links between concept pairs of a

**Figure 4:** Fine-tuning the BERT\textsubscript{BASE} model with concept pairs to obtain BERT\textsubscript{BASE+CLF} model.
We randomly masked out two tokens “##mon” and “##tis” over other tokens in the vocabulary. The BERT Base model computed the probabilities for Class 0 and Class 1, and recorded the result as a 2 x 1 vector. The classifier reported the class label with the higher probability. The error between the true label and the label predicted by the model was back-propagated through the model to improve the network’s parameters. The obtained model is denoted as BERT Base-CLF (CLF = classifier), the model after Fine-tuning. For this we used the default model hyperparameters in pre-trained BERT Base, with the exception of the sequence length (=128), batch size (=64), learning rate (=2e-5), and number of training epochs (=3).

Pre-training and fine-tuning strategy

Data Preparation: In our setup of unsupervised Pre-training, BERT Base is not trained for a specific task, but the purpose is integrating medical knowledge into its representation. This is done by training with non-task related samples taken from SNOMED CT. BERT was originally trained with millions of documents that are composed of sentences. Similarly, we generated a list of terminology-oriented documents by creating one “document” per focus concept F (Figure 5), with related concept(s) as “sentences” of such documents. The ID for a document is the corresponding SNOMED CT concept ID. The content of this document consists of the concepts that are hierarchically related to F (Figure 5(a)). Specifically, we chose F’s parents (targets of IS-A links from F), F itself, and its children (sources of IS-A links to F). Thus, the whole document text is: Parent(s) – Focus concept – Child(ren) (Figure 5(b)). The concept groups are separated into lines, e.g., “(Parents) finding of abnormal level of heavy metals in blood, finding of trace element level – NEW LINE– (Focus concept) blood copper abnormal –NEW LINE– (Children) raised blood copper level, serum copper level abnormal” is the document for the focus concept blood copper abnormal. This construction is based on the idea that a concept is the topic of a document and that the closely related concepts are descriptions of the meaning of this concept in the terminology hierarchy. To feed sentences into the BERT Base model, all the documents are concatenated in one text file, separated by empty lines.

Pre-training the BERT Base model: To utilize BERT’s powerful language representation, we started with BERT Base and embedded terminology knowledge with new training data. BERT was originally trained for two unsupervised prediction tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) on an arbitrary text corpus. We adopted the same two training tasks and objective with concept-based documents from SNOMED CT. In the MLM phase, the training objective is to predict only the masked words. We randomly masked out 15% of the words across all the concept-based documents, and then trained the complete BERT Base model to output the masked words. In the NSP phase, the objective is to learn relationships between concepts: Given two concepts A and B, is B a child of A, or not (Figure 6). We extracted two “sentences” Colitis and Phlegmonous colitis from the document for the focus concept Colitis. After preprocessing these two concepts (treated as two “sentences”) as shown in the middle level, we masked out two tokens – “##mon” and “##tis”. The BERT Base model was trained to raise the probabilities of two correct tokens “##mon” and “##tis” over other tokens in the vocabulary. In addition, as Phlegmonous colitis IS-A Colitis, the BERT Base model was also trained to output the correct classification label “IsNext.” The obtained model
is denoted as $\text{BERT}_{\text{BASE+SNO}}$ (SNO=SNOMED CT). Then we applied the Fine-tuning process (see above) to the $\text{BERT}_{\text{BASE+SNO}}$ model to get the $\text{BERT}_{\text{BASE+SNO+CLF}}$ model.

The training parameters used for Pre-training are as follows: batch size = 64, sequence length =128, training steps = 15,000 for Procedure and 200,000 for Clinical Finding, learning rate = $2 \times 10^{-5}$, dropout rate = 0.1, and activation function = gelu (Gaussian error linear unit).

**Figure 6:** Pre-training the $\text{BERT}_{\text{BASE}}$ model with concept-based documents to obtain $\text{BERT}_{\text{BASE+SNO}}$ model. FFNN is short for Feedforward neural network.

**Figure 7:** Data flow for testing the trained $\text{BERT}_{\text{BASE+CLF}}$ or $\text{BERT}_{\text{BASE+SNO+CLF}}$ models with unseen data.
Testing the prediction on new release data

To evaluate the BERT\textsuperscript{BASE+CLF} and the BERT\textsuperscript{BASE+SNO+CLF} models on previously unseen data, we created separate test tasks, using new concepts from the Procedure and the Clinical finding hierarchy of the January 2018 release (Figure 7). This description will focus on the Procedure hierarchy. For each new concept that was added to the Procedure hierarchy in this release, we extracted it and its parents as positive sample pairs. For example, Local excision of lesion of kidney has two parents Local excision and Excision of lesion of kidney. The corresponding positive testing samples are “Local excision (t) Local excision of lesion of kidney” and “Excision of lesion of kidney (t) Local excision of lesion of kidney” with the true Class label = 1. For the negative sample, we paired each new concept with a randomly chosen concept from the other new concepts’ parents. For example, we randomly select Ultrasonography of left lower limb, which is the parent of Ultrasonography of left knee region, and paired it with Local excision of lesion of kidney to form the instance “Ultrasonography of left lower limb (t) Local excision of lesion of kidney” with the label = 0.

The concept pairs were randomly arranged into a sequence and sent to the trained BERT\textsuperscript{BASE+CLF} and BERT\textsuperscript{BASE+SNO+CLF} models. The tested model processed each input pair, using the weights that it had learned before, returning a class label (0 or 1) as prediction result. For negative samples, label 0 is correct, indicating that there is no IS-A link between these two concepts in the new SNOMED CT release. Label 1 is correct for positive samples, indicating the existence of an IS-A link. The predicted result labels were compared with the true labels to calculate prediction accuracy in terms of Precision, Recall, and F1 score.

Results

We first report the prediction results of the Fine-tuning model and Pre-training & Fine-tuning model with samples extracted from the Procedure hierarchy of the SNOMED CT 2018 January release, with 15,000 training steps. The Precision, Recall, and F1 scores for ten tests are presented in Table 1. For the Procedure hierarchy, the model was tested against 3,908 pairs (1,954 positives and 1,954 negatives). For example, in Test 7 for IS-A classification, the Precision is 0.69, Recall is 0.98, and F1 score is 0.81 for Fine-tuning. When adding Pre-training, Precision is 0.73, Recall is 0.98, and F1 score is 0.84. The F1 score improved by about 3.7%. Similarly, for Non-IS-A tests, the F1 score increased from 0.71 to 0.77, an 8.5% improvement. On average, by adding Pre-training, there are 6.3% (from 0.80 to 0.85) and 14.5% (from 0.69 to 0.79) improvements of F1 for IS-A and Non-IS-A classifications, respectively.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>IS-A Classification</th>
<th>Non-IS-A Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fine-tuning</td>
<td>Pre-training &amp; Fine-tuning</td>
</tr>
<tr>
<td>No.</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>1</td>
<td>0.67</td>
<td>0.98</td>
</tr>
<tr>
<td>2</td>
<td>0.66</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>0.70</td>
<td>0.98</td>
</tr>
<tr>
<td>4</td>
<td>0.66</td>
<td>0.98</td>
</tr>
<tr>
<td>5</td>
<td>0.69</td>
<td>0.98</td>
</tr>
<tr>
<td>6</td>
<td>0.71</td>
<td>0.97</td>
</tr>
<tr>
<td>7</td>
<td>0.69</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td>0.67</td>
<td>0.98</td>
</tr>
<tr>
<td>9</td>
<td>0.69</td>
<td>0.98</td>
</tr>
<tr>
<td>10</td>
<td>0.66</td>
<td>0.98</td>
</tr>
<tr>
<td>Average</td>
<td>0.68</td>
<td>0.98</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Due to space limitations, we only report the summary of the Precision, Recall, and F1 scores from ten tests for the Procedure hierarchy with 10,000 training steps (Table 2). In each test, the model was tested against 3,908 pairs (1,954 positives and 1,954 negatives). On average, by adding Pre-training to Fine-tuning, the improvements are 3.75% (from 0.80 to 0.83) and 10.1% (from 0.69 to 0.76) for IS-A and Non-IS-A classifications, respectively.
Table 2. Precision, Recall, and F1 score for ten tests of Procedure hierarchy (training steps = 10,000).

<table>
<thead>
<tr>
<th>Procedure</th>
<th>IS-A Classification</th>
<th>Non-IS-A Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fine-tuning</td>
<td>Pre-training &amp; Fine-tuning</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Average</td>
<td>0.68</td>
<td>0.98</td>
</tr>
<tr>
<td>Max</td>
<td>0.71</td>
<td>0.98</td>
</tr>
<tr>
<td>Min</td>
<td>0.66</td>
<td>0.97</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

For the Clinical finding hierarchy, the summary of ten tests results with training steps = 200,000 is reported in Table 3. In each test, the model was tested against 8,574 pairs (4,287 positives and 4,287 negatives). On average, by adding Pre-training, the improvements are 7.5% (from 0.80 to 0.86) and 15.3% (from 0.72 to 0.83) for IS-A and Non-IS-A classifications, respectively.

Table 3. Precision, Recall, and F1 score for ten tests of Clinical Finding hierarchy (training steps = 200,000).

<table>
<thead>
<tr>
<th>Clinical Finding</th>
<th>IS-A Classification</th>
<th>Non-IS-A Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fine-tuning</td>
<td>Pre-training &amp; Fine-tuning</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Average</td>
<td>0.70</td>
<td>0.94</td>
</tr>
<tr>
<td>Max</td>
<td>0.72</td>
<td>0.94</td>
</tr>
<tr>
<td>Min</td>
<td>0.69</td>
<td>0.93</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4. Prediction results of two models on five IS-A & five non-IS-A examples from Clinical Finding hierarchy.

<table>
<thead>
<tr>
<th>Index</th>
<th>Test Concept</th>
<th>New Concept</th>
<th>True Label</th>
<th>Fine-tuning</th>
<th>Pre-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Injury of trachea</td>
<td>Crushing injury of trachea</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Arthropathy of knee joint</td>
<td>Aseptic necrosis of right lateral femoral condyle</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Lesion of neck</td>
<td>Stenosis of right vertebral artery</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Persistent pain following procedure</td>
<td>Chronic pain following radiotherapy</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Joint injury</td>
<td>Traumatic rupture of ligament of wrist</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Bursitis of shoulder</td>
<td>Injury of toenail</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Atherosclerosis of artery</td>
<td>Crushing injury of trachea</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>Finding of employment status</td>
<td>Social isolation in parenthood</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>Soft tissue injury</td>
<td>Injury of bilateral optic tracts</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>Injury of wrist</td>
<td>Injury of peripheral nerve of abdomen</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Regarding the prediction of IS-A links for new concepts, we show ten examples of our two models’ prediction results (Table 4) for ten pairs for which the second concept was newly added to SNOMED CT’s Clinical finding hierarchy in the 2018 January release. For each test, we paired one Test Concept with one New Concept as one test instance,
then we let the model predict IS-A links between them. For instance, for Example 2, we chose Arthropathy of knee joint as the first concept and paired it with Aseptic necrosis of right lateral femoral condyle. Then the task became to predict whether there is an IS-A link between the two concepts. Both the Fine-tuning and Pre-training & Fine-tuning models returned the correct label (=1). For Example 4, the Fine-tuning model is wrong, and the combined model is correct that Chronic pain following radiotherapy IS-A Persistent pain following procedure. Both models are wrong about Injury of bilateral optic tracts, because it is not a Soft tissue injury (Example 9).

**Discussion**

In this paper we set out to investigate whether the “next sentence prediction” capability of BERT can be fine-tuned to verify the parent(s) of new concepts added to a terminology. Such a capability can be utilized in automatic enrichment of terminologies. The results for the Procedure and Clinical finding hierarchy confirm that our technique, which utilizes this capability of BERT, is indeed able to verify the IS-A relationships from new concepts to their parents with a 0.80 F1 average value. When enhancing Fine-tuning of BERT with Pre-training the average F1 score grows to 0.85 (0.86).

However, looking at the details we see that the recall to identify IS-A relationships is very high (0.98) while the precision is only 0.74. The outcome for identifying the Non-IS-A pairs of concepts is the opposite, with high (0.97) precision and low (0.66) recall, yielding an F1 value of 0.79. These results indicate that our technique verifies almost all the IS-A relationships, but wrongly identifies some Non-IS-A pairs as having IS-A relationships.

Thus, the challenge for future research is to improve the precision. In previous research [11, 21], we have shown that using summarization techniques of terminologies [22] can improve the training of ML techniques to better distinguish between IS-A relationships and pairs that are not connected by IS-A relationships. In future research, we will investigate whether utilizing summarization techniques can improve the precision and thus the F1 value.

Another issue is the number of training steps required for Pre-training. For the Procedure hierarchy 15,000 training steps provided the best results for enhancing the process by Pre-training. The enhancement was about double than for 10,000 training steps. Experiments with 20,000 and 25,000 training steps showed a leveling off of F1 at 20,000 and a lower F1 value at 25,000. However more research is needed to confirm this behavior beyond 15,000 training steps. For the Clinical finding hierarchy, which is about twice the size of the Procedure hierarchy, 200,000 training steps were required to obtain the same enhancement as for the Procedure hierarchy, which had required just 15,000 training steps. Hence, much more pre-training steps are required when the hierarchy is larger. The two transfer learning models using BERT are superior to our previously proposed CNN model [11], which was trained with the whole SNOMED CT consisting of 473,756 concepts. That CNN model achieved an average F1 score of 0.70, and could only verify IS-A links for new concepts with multiple parents. As a consequence of this observation, it is preferable to perform automatic enrichment for each hierarchy of SNOMED CT separately, rather than for the whole SNOMED CT. Another reason for enriching each hierarchy separately is that the content of each hierarchy is different, and ML is likely more effective by modularizing the learning into uniform hierarchies than learning a large, non-uniform body of knowledge in one process.

In the BERT\textsubscript{BASE} model, features are more generic/linguistic in the early network layers and more dataset-specific in the later layers. Thus, fine-tuning is normally inexpensive, because one only needs to modify the later layers or train one or two task-specific layers on top. All of the results in the paper can be replicated in at most 3 to 5 hours on a single GPU, starting from the same pre-trained BERT\textsubscript{BASE} model. Pre-training is more expensive than Fine-tuning. For example, it took about four days to run 200,000 steps to pre-train the model with the Clinical finding hierarchy data on a single GPU. However, this is a one-time procedure for each hierarchy. We plan to release the two pre-trained models of this paper for future research work, to save other researchers the effort and time to pre-train their own models from scratch.

**Limitations:** The BERT\textsubscript{BASE} model was trained with the concatenation of the BooksCorpus (800M words) [20] and English Wikipedia (2,500M words). It employs the WordPiece embeddings [23] with a 30,522 tokens vocabulary, which does not include most medical terms. Thus, medical terms that are not in the WordPiece vocabulary are split into word pieces denoted by ##. For example, “adenoid” is split into “aden” and “##oid.” The lack of medical terms in BERT’s vocabulary limited its applicability to support insertion of new concepts into a medical terminology such as SNOMED CT, and would probably impair other NLP tasks within the medical domain. In future work, we will expand the vocabulary to include common medical terms selected from terminologies such as SNOMED CT. We will investigate whether pre-training BERT with medical terms can help improve its performance in some common NLP tasks in the medical domain, such as tagging and named entity recognition in EHRs.
Conclusion
We have shown that one can fine-tune the BERT model and obtain an effective technique for correctly placing new concepts in the right positions of a terminology. Furthermore, by pre-training BERT with SNOMED CT content, we improved the precision while preserving the high recall and thus we improved the F1 value.

Acknowledgments: Research reported in this publication was partially supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References
Multivariate Analysis of Physicians’ Practicing Behaviors in an Urgent Care Telemedicine Intervention

Songzi Liu, BA1, Barbara Edson, RN, MHA, MBA2, Robert Gianforcaro, DO, FAAFP2, Khairat Saif, PhD, MPH, FAMIA3
1The School of Information and Library Science, University of North Carolina, Chapel Hill, NC, USA; 2Virtual Care Center, UNC Healthcare, Chapel Hill, NC, USA; 3School of Nursing, University of North Carolina, Chapel Hill, NC, 27514, USA

Abstract
When assessing the characteristics and performance of telemedicine interventions, most studies followed a patient-centric approach, leaving the telemedicine providers’ role out of consideration. As a result, little was known about the demographics and prescription pattern of telemedicine physicians, the knowledge of which is integral to a holistic evaluation of the virtual delivery of accountable care. To fill this gap, our study explored how physicians’ traits and encounter-specific characteristics correlate with prescription outcomes, using multivariate analyses. Significant inter-physician variation in prescription behaviors was observed and analyzed in sub-groups. The average Virtual Urgent Care physician’s prescription likelihood was 69% with a mean prescription count of 0.98; male physicians and primary care providers tended to prescribe both more often and with a greater number of medications. This study called attention to the quality and reproducibility of telemedicine providers’ prescription decision and warned the likely absence of well-defined practice guidelines for delivering virtual care.

Introduction
Telemedicine facilitates patient-provider interaction and the delivery of clinical consultation via information technology1. While most telemedicine research focuses primarily on patient’s characteristics, clinical effectiveness, and technical outcomes3,4, limited number of studies investigated the physicians’ prescribing behaviors in a virtual setting.

Telemedicine Providers’ Prescribing Patterns
In the past, researchers were primarily concerned with telemedicine providers’ performance in terms of diagnostic accuracy, obtained by analyzing medical images in certain specialties such as radiology, dermatology, ultrasound, and pathology6. However, this approach was not applicable to common acute illnesses, which made up the majority of current commercial virtual visits7. Only a handful of studies shed light on the demographics and practicing behaviors of telemedicine providers. Previously, researchers studied the difference between in-person and telemedicine providers’ prescribing patterns, concluding that provider type was a significant influence in determining the count of medication prescribed per consultation8. Another study investigated the prescription behaviors of General Practitioners in a Danish out-of-hours primary care telephone consultation practice. The authors further explored the association between various patients and physicians’ characteristics and prescription frequency, concluding that one in five phone consultations ended with a prescription; the prescription likelihood was the highest from 4 to 8 p.m. on weekdays and during the early morning on weekends9.

These findings, to some extent, complemented the previous research on inter-physician variation in prescribing practice in the traditional, face-to-face setting, which generally suggested a positive correlation between the providers’ personal characteristics and prescription outcomes10,11. However, the presence of a significant knowledge gap in telemedicine physicians precluded a systematic, direct comparison between the in-person and virtual physician populations. Therefore, the knowledge of telemedicine providers will offer valuable insights on the patient-physician relationship in the virtual environment, which has become an urgent matter as telemedicine gained more acceptance in recent years.

Virtual Urgent Care
This study is uniquely important in a sense that it investigated a rarely-studied type of telemedicine service—a Virtual Urgent Clinic (VUC) telemedicine platform, providing virtual medical consultation to the general population in North Carolina 24-hours a day, 7-day a week. More importantly, the VUC physicians come from four areas of specialties, which enabled between-provider-type comparison. Encounters characteristics, such as time of day, day of the week,
modality, and encounter duration were analyzed to provide a comprehensive understanding of telemedicine providers’ prescribing behaviors.

**Objective**
The goals of this paper were to further understand the characteristics of VUC physicians and to identify pattern and variation in their prescribing behaviors, using a combined method of descriptive and multivariate analysis. In addition, we discussed the implications of the providers’ behaviors in improving the delivery of accessible and accountable virtual care.

**Methods and Materials**

**Program Description**
We investigated the first 11-month operation data from the 24/7 virtual urgent care service in 2018. The VUC provides patient care across North Carolina with synchronous, virtual consultation for illnesses of acute nature—around-the-clock and 7 days a week—enabled by a third-party telemedicine platform.

Registration is mandatory for all patients in order to access the VUC service. Upon registration, the users are required to provide both demographic data (including age, gender, zip codes of residence, and status of dependency) and detailed accounts of medical history such as allergies, drug sensitivities, and family health history. At the point of scheduling a consultation, the patients are presented with a comprehensive list of conditions which could not be treated via the platform. The patients specify the reason for the visit from a pre-identified list of conditions and choose a provider per one’s preferences and availability. Moreover, the patient will choose either a phone call or video conference as their preferred encounter medium. Additional instructions are provided in case where the required software has not been appropriately installed. During the consultation, the VUC provider evaluates the patient symptoms and document the diagnosis in the standard ICD-9-CM Diagnosis Code; in situations which result in a prescription of any medications, the physician will send an electronic prescription to the patient’s pre-assigned pharmacy. The VUC service cost is a flat fee regardless of consultation types or duration. The study was approved by the Institutional Review Board at the University of North Carolina-Chapel Hill.

**Study Population**
VUC delivers urgent medical care to residents of North Carolina above the age of two-year-old. In addition, the patients will need access to a smartphone or a computer with the required software installed. VUC serves patients regardless of their demographic traits, insurance coverage, or other status.

All VUC third-party providers were board-certified physicians. We collected data of 26 VUC providers, including 9 Family Medicine, 8 Internal Medicine, 6 Emergency Medicine, 2 Pediatrics, and 1 General Practice physician. The sole General Practice provider was excluded from the analysis due to its inability to represent the General Practice physician population, which is likely to compromise the overall quality of the study. We decided to focus on two main types of providers: Primary Care—which encompassed Family Medicine, Internal Medicine, and Pediatrics providers—and Emergency Medicine.

**Dataset and Variables**
For analysis purpose, 141 incomplete and test-user encounters were excluded. Eight independent variables from two categories, the physicians’ characteristics and the encounter-specific features, were extracted and coded.

The provider’s gender, enunciation date, and medical specialty were identified through looking up their National Provider Identifiers(NPI), which was included in the original dataset, on the NPPES NPI Registry website. The procedure was aided by applying generic web-scraping algorithm in Python. The variable Year of Practice was constructed by calculating the difference between the physician’s year of enunciation and the current year (2019) and was subsequently categorized into over 10 years and under 10 years. Additionally, we counted the number of encounters completed by each provider and constructed Practice Frequency, indicating an individual’s degree of involvement in practicing telemedicine; likewise, the variable was divided into three categories—over 100 times, 51-100 times, and 11-50 times. Encounter-specific features were also recorded in the original VUC dataset, including Time of Day, Day of Week, Encounter Modality, and Encounter Duration. The features were classified into subgroups for the purpose of evaluation, as shown in Table 1. For additional inter-physician prescription variation analysis, we divided the dataset by the top five popular diagnosis codes: Sinusitis, Urinary Tract Infection, Pharyngitis, and Bronchitis).
Outcome Variables

The primary outcomes of interest were prescription likelihood and prescription count. Prescription Likelihood denoted the possibility of whether an encounter ended with medication prescribed; it was calculated by averaging the binary prescription outcome of encounters under a certain criterion, where 0 indicated no prescription and 1 indicated at least one count of medication. The variable Prescription Count recorded the exact number of medications prescribed per session on a range of 0 to 6.

Statistical Analysis

Descriptive analysis was conducted on the test population, giving the mean prescription likelihood and count by physicians and encounter characteristics. For the count variable Prescription Count, we applied the Kruskal-Wallis test to determine if the means in a group were significantly different from one another. The non-parametric Kruskal-Wallis test was chosen over the One-way ANOVA since the former does not assume the normality condition\textsuperscript{13}. For the nominal variable Prescription Likelihood, Pearson Chi-squared test was used for assessing the in-group difference. In addition, the top five most popular diagnoses were extracted and tested individually for the presence of inter-physician difference in prescribing outcome using Chi-squared test.

Variable Selection and Modeling

To remove the unnecessary predictors, a Stepwise Regression was run with all eight independent variables. Five predictors were kept for the final modelling while four were excluded. Doing so help reducing the noise which may arise from the collinearity among the redundant predictors.

Subsequently, a Multivariate Logistic Regression model was constructed to investigate the association between Prescription Likelihood and the five independent variables; log ratios and p-values were calculated. In addition, a Poisson Regression was built for estimating the impact of the independent variables on Prescription Count. Based on Cameron and Trivedi’s recommendation in their 2009 study, we calculated the robust standard errors for the parameter estimates in order to compensate for mild violation of the equal variance assumption\textsuperscript{14}.

Data Collection and Cleaning

Encounter data was automatically scrapped from the service’s website and stored in a secure SAP© Business Objects (BO) enterprise system. The dataset included physicians’ unique identifier, diagnosis code, prescription count, and encounters characteristics such as time of day, day of the week, duration, and modality information. The dataset was cleaned and preprocessed with Excel and OpenRefine. The statistical analyses and models were generated using R (R Foundation for Statistical Computing, 2014). The data visualization was made with Tableau Desktop 2019.

Results

This study analyzed a total of 1,217 encounters completed by 25 physicians, collected during the 11-month period between February 8\textsuperscript{th} and November 29\textsuperscript{th}, 2018. On average, 69\% of VUC encounters ended with at least one medication prescribed, with a mean prescription count of 0.98 per encounter.

Inter-physician Variation in Prescribing Behavior

To evaluate the presence of inter-physician variation in prescribing behavior, we assigned each provider a unique identifier, Provider ID (ranging from 1 to 25), which was subsequently tested as a predictor of prescription outcome. A Pearson Chi-squared test of Provider ID and Prescription Likelihood yielded a p-value of 0, indicating an extremely strong correlation between the individual provider’s accumulated characteristics and prescription likelihood. In other words, the 25 physicians prescribed in significantly different manners. Figure 1. visualized the frequency of prescribing a certain number of medications (0, 1, 2, and 3+) as a percentage of a provider’s total encounters.

In encounters resulting in prescription, one count of medication was the most common case. However, some physicians showed a tendency of prescribing multiple medications per encounter; for instance, provider 6 prescribed more than one medication per session about 90\% of the time and frequently prescribed greater than three counts. On the other hand, some providers showed reluctance towards prescribing at all (Provider 4, 12, 23), Figure 1.
Figure 1. Individual Difference among VUC Physicians’ Prescription Counts.

Descriptive Analysis

Providers Characteristics

Among the 1,217 encounters, 71.5% were completed by male physicians, who, on average, prescribed 24% more frequent than the female counterpart, Table 1. Of all consultations, 84.72% were completed by providers with over ten years of experience. Physicians with less than ten years of practice tended to prescribe a significantly greater number of drugs per encounter; however, they were not more likely to prescribe than their counterpart. Furthermore, Primary Care providers completed the most number of consultations (68.45%), while Emergency Medicine contributed the remaining 31.55%. The Chi-squared test results of both outcome metrics indicated significant difference between provider types. Lastly, 40.76% of the consultations were conducted by physicians who practiced less often (11-50 times), while physicians who had completed over 100 sessions contributed 37.14%. The between-group differences were significant in both outcome measures.

Encounters Characteristics

Of all encounters, 39.77% took place in the morning (6 a.m. - 12 p.m.); afternoon (12 p.m. - 5 p.m.) and evening sessions (5 p.m. - 10 p.m. and 10 p.m. - 6 a.m. combined) composited around 30% of total encounters respectively. Physicians were more likely to prescribe during morning hours, where 73% of sessions ended with drugs prescribed; sessions happened after 10 p.m. had the lowest prescription likelihood of 61%. Likewise, morning sessions had the highest prescription count (1.04), which is 73.56% higher than the average count of late night sessions (0.84). Additionally, approximately 7 out of 10 encounters lasted between 1 and 5 minutes. Durations of less than 1 minute and over 10 minutes were rare, with less than 2% and 6% rates of occurrence respectively. Furthermore, about 3 out of 10 encounters happened during the weekend, the prescription likelihood and count of which were both significantly lower than those of the Weekday encounters. In terms of telemedicine modality, Phone (88%) was predominantly more popular than Video (12%); the difference in prescription outcomes between the two, however, was not significant.
**Table 1.** Average Prescription Likelihood and Prescription Count by Physicians and Encounter characteristics.

<table>
<thead>
<tr>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (%)</td>
<td>likelihod</td>
<td>Presc. Likelihood</td>
</tr>
<tr>
<td>All Encounters (N)</td>
<td>1,217(100%)</td>
<td>0.69</td>
</tr>
</tbody>
</table>

**Physicians Characteristics**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>347(28.51%)</td>
<td>0.52</td>
<td>0.05</td>
</tr>
<tr>
<td>Male</td>
<td>870(71.49%)</td>
<td>0.76</td>
<td>0.03</td>
</tr>
</tbody>
</table>

**Year of Practice**

<table>
<thead>
<tr>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over 10 years</td>
<td>1031 (84.72%)</td>
<td>0.68</td>
</tr>
<tr>
<td>Under 10 years</td>
<td>186 (15.28%)</td>
<td>0.72</td>
</tr>
</tbody>
</table>

**Specialty**

<table>
<thead>
<tr>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emergency Medicine</td>
<td>384(31.55%)</td>
<td>0.64</td>
</tr>
<tr>
<td>Primary Care</td>
<td>833(68.45%)</td>
<td>0.71</td>
</tr>
</tbody>
</table>

**Prescription Frequency**

<table>
<thead>
<tr>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-50 times</td>
<td>496(40.76%)</td>
<td>0.71</td>
</tr>
<tr>
<td>51-100 Times</td>
<td>269(22.11%)</td>
<td>0.56</td>
</tr>
<tr>
<td>Over 100 Times</td>
<td>452(37.14%)</td>
<td>0.75</td>
</tr>
</tbody>
</table>

**Encounter Characteristics**

<table>
<thead>
<tr>
<th>Time of Day</th>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 a.m. - 12 p.m.</td>
<td>484(39.77%)</td>
<td>0.73</td>
<td>0.04</td>
</tr>
<tr>
<td>12 p.m. - 5 p.m.</td>
<td>364(29.91%)</td>
<td>0.65</td>
<td>0.05</td>
</tr>
<tr>
<td>5 p.m. - 10 p.m.</td>
<td>272(22.35%)</td>
<td>0.69</td>
<td>0.06</td>
</tr>
<tr>
<td>10 p.m. - 6 a.m.</td>
<td>97(7.97%)</td>
<td>0.61</td>
<td>0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Day of week</th>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekday</td>
<td>896(73.62%)</td>
<td>0.71</td>
<td>0.03</td>
</tr>
<tr>
<td>Weekend</td>
<td>321(26.38%)</td>
<td>0.63</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Duration of encounter</th>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1 Min</td>
<td>21(1.73%)</td>
<td>0.76</td>
<td>0.20</td>
</tr>
<tr>
<td>1-5 Min</td>
<td>839(68.94%)</td>
<td>0.70</td>
<td>0.03</td>
</tr>
<tr>
<td>6-10 Min</td>
<td>289(23.75%)</td>
<td>0.64</td>
<td>0.06</td>
</tr>
<tr>
<td>&gt;10 Min</td>
<td>70(5.59)</td>
<td>0.69</td>
<td>0.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Population</th>
<th>Presc. likelihood (95% CI)</th>
<th>Presc. Count (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>1071(88%)</td>
<td>0.69</td>
<td>0.03</td>
</tr>
<tr>
<td>Video</td>
<td>146(12)</td>
<td>0.68</td>
<td>0.08</td>
</tr>
</tbody>
</table>

*a* denotes Kruskal-Wallis rank sum test p-value <0.05 for characteristics—prescription count.

**β** denotes Chi-squared test p-value <0.05 for characteristics—prescription likelihood.
**Prescription Outcomes Analysis**

**Prescription Rate**

To further understand whether a certain physician or encounter characteristic could potentially affect the likelihood of prescription at a given consultation encounter, a Multivariate Logistic Regression was run to determine both the significance and magnitude of each independent variable’s predictive power, as shown in Table 2.

The resulting test statistics indicated that Provider Gender and Specialty was the most significant predictors, with p values of less than 0.05; Time of Day and Day of Week were less strong but still marginally significant, with p-values of less than 0.1. Odds ratio were subsequently calculated. For instance, all else held equal, male providers were 2.1 times more likely to prescribe than their female counterpart. Primary Care providers were 42% more likely to prescribe than the EM specialists. On the other hand, a Weekend encounter, completed by a physician with the same set of characteristics and shared the identical encounter features as its Weekday counterpart, was 78% less likely to end with medication prescribed. In addition, consultations happened in the morning hours were about twice more likely to get prescription as compared to the sessions taken place after 10 p.m..

**Prescription Count**

A Poisson Regression with Prescription Count and the set of 5 variables was performed. To ease the interpretation, we transformed the resulting log-odds estimate into incidence rate ratio, Table 2.

Provider Gender and Specialty turned out to be significant predictors, with p-values of less than 0.01 and 0.1 respectively. All else features held equal, male physicians prescribe 48% unit more than the female. Under the same encounter settings, Primary Care providers prescribe 15% more than the EM specialists of their own gender and practice frequency.

Table 2. Multivariate Logistic Regression and Poisson Regression.

<table>
<thead>
<tr>
<th>Independent Var.</th>
<th>Odds Ratio</th>
<th>p value</th>
<th>Incidence Rate Estimate</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.78</td>
<td>0.36</td>
<td>0.58***</td>
<td>0</td>
</tr>
<tr>
<td>Time of Day</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 p.m. - 5 p.m.</td>
<td>1.04</td>
<td>0.86</td>
<td>1.09</td>
<td>0.49</td>
</tr>
<tr>
<td>5 p.m. - 10 p.m.</td>
<td>1.31</td>
<td>0.29</td>
<td>1.15</td>
<td>0.23</td>
</tr>
<tr>
<td>6 a.m. - 12 p.m.</td>
<td>1.59*</td>
<td>0.06</td>
<td>1.26</td>
<td>0.05</td>
</tr>
<tr>
<td>Day of Week</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weekend</td>
<td>0.78*</td>
<td>0.08</td>
<td>0.96</td>
<td>0.51</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3.10***</td>
<td>0.00</td>
<td>1.48***</td>
<td>0</td>
</tr>
<tr>
<td>Specialty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary Care</td>
<td>1.42**</td>
<td>0.02</td>
<td>1.15**</td>
<td>0.02</td>
</tr>
<tr>
<td>Practice Frequency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51-100 times</td>
<td>0.48</td>
<td>0.38</td>
<td>0.97</td>
<td>0.68</td>
</tr>
<tr>
<td>Over 100 times</td>
<td>1.15</td>
<td>0.40</td>
<td>0.99</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Akaike Inf. Crit. 1,355.739
Null deviance: 1150.3 on 1216 degrees of freedom
Residual deviance: 1103.1 on 1208 degrees of freedom
AIC: 2991.5
Note: *p<0.1; **p<0.05; ***p<0.01

1144
Prescription Outcome Analysis for the Top 5 Diagnosis

The five most popular diagnosis (accumulated n=463, 38.04% of total encounters) were individually tested for inter-physician variation in prescribing outcome. The statistics were obtained using Chi-squared test, as shown in Table 3.

Different physicians prescribed in significantly different manner when treating Sinusitis, Urinary Tract Infection, Pharyngitis, and Upper Respiratory Infection (p value<0.05). However, the inter-physician variation was not observed in treating Bronchitis. Bronchitis patients were also the most likely to get prescription (prescription likelihood =85.7%). In case of Upper Respiratory Infection, only 63.8% encounters ended with any drug prescribed.

Table 3. Prescribing Patterns in the Top 5 Diagnosis

<table>
<thead>
<tr>
<th>Diagnosis Name (ICD-10-CM Code)</th>
<th>Count</th>
<th>Presc. Likelihood</th>
<th>95% CI</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinusitis (J01.90)</td>
<td>198</td>
<td>0.833</td>
<td>0.781-0.886</td>
<td>0.003***</td>
</tr>
<tr>
<td>Urinary Tract Infection (N39.0)</td>
<td>86</td>
<td>0.80</td>
<td>0.716-0.888</td>
<td>0.02**</td>
</tr>
<tr>
<td>Pharyngitis (J02.0)</td>
<td>65</td>
<td>0.708</td>
<td>0.594-0.821</td>
<td></td>
</tr>
<tr>
<td>Upper Respiratory Infection (J06.9)</td>
<td>58</td>
<td>0.638</td>
<td>0.510-0.765</td>
<td>0.05**</td>
</tr>
<tr>
<td>Bronchitis (J20)</td>
<td>56</td>
<td>0.857</td>
<td>0.763-0.952</td>
<td>0.37</td>
</tr>
<tr>
<td>Total</td>
<td>463</td>
<td>0.788</td>
<td>0.751-0.826</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: p-value was calculated using Chi-squared Test.
* p<0.1; ** p<0.05; *** p<0.01

Discussion

Main Findings

We found that virtual care providers were predominantly male primary care physicians with over ten years of experience. Male providers displayed a tendency toward prescribing a greater number of medications at any given consultation. In addition, the difference between Primary Care and Emergency Medicine physicians were evident in both outcome measures, with the latter being more reluctant toward prescribing and suggesting a higher count of drugs. A marginally significant correlation existed between temporal factors (Time of Day, Day of the Week) and prescription outcomes. On the other hand, the impact of telemedicine modality and duration of encounter were not evident in affecting the physician’s prescribing behaviors. Inter-physician variation in prescription outcome was the most present in treating diseases of the respiratory system (ICD-10-CM Diagnosis Code J00-J99), which were also the most popular VUC diagnoses.

Previous Study

Telemedicine Interventions

In terms of prescription outcomes, the variations were considerable: The average prescription likelihood of VUC encounters was 0.69, which was significantly higher than the rate of 0.19 in a similar study of a telephone consultation intervention in Denmark. On the other hand, VUC’s mean prescription count was 0.98, which was 29.6% lower than the 1.27 average in another e-visits intervention. However, the scarcity of comparable literature prevented us from understanding the underlying pattern of such pronounced differences among individual telemedicine interventions. These great variances alarmed the potential absence of common practice guidelines in the virtual setting, raising questions on the fairness and quality of telemedicine delivery.

In-Person Urgent Care Visit

According to a national survey conducted with 436 in-person Urgent Care Centers in the United States, primary care providers managed 72% of total encounters, while emergency medicine physicians and non-primary care specialists contributed 28% and 20% respectively. Our study indicated a similar composition of physician’s specialty, with a 2:1 ratio between primary care and emergency medicine physicians. However, there was an absence of non-primary care specialists among the telemedicine providers. We suspected that the VUC protocol, designed specifically for the
assessment of common acute illnesses via voice or video conferencing, prohibited advanced patient-physician interaction that might be required in the specialists’ consultation.

Time and day of urgent care visit were significantly different between in-person and virtual settings. One study sampled 387,746 records of acute care visit in the U.S., collected between 1997 and 2010, suggesting that 95 percent of visits to office-based primary care providers happened on weekdays. On the other hand, three out of ten VUC encounters took place on weekends. The virtual urgent care demonstrated potential in providing more timely and accessible medical service due to the around-the-clock availability of telemedicine physicians. In other words, telemedicine service could complement the office-based delivery of primary care, ameliorating the staffing problem faced by many healthcare organizations nowadays.

**In-person Physician**

Our study agreed with a number of researches on that male physicians tend to be high-prescribers. One study concluded that physicians who had more practice days and saw more patients per day would also prescribe greater number of drugs. While our findings indicated significant between-group differences among telemedicine physicians with varied frequency of practice, the variable itself was not an effective predictor of both prescribing outcomes in a virtual setting.

**Strengths and Limitations**

This study was the first to investigate the characteristics and prescription behaviors of virtual urgent care telemedicine providers, with first-hand data collected from a Southeastern commercial telemedicine platform. In addition, we searched for literature on office-based urgent care centers in the United States and compared the operational statistics with the VUC dataset. However, the inter-physician variation in prescribing styles suggested that one’s prescribing decision might be prone to personal biases, which we were not able to validate in this paper.

In addition, this paper did not directly compare telemedicine physicians with in-person physician due to the lack of access to comparable data collected from in-person urgent care settings. We were only able to compare our findings with similar telemedicine evaluation studies. The implications of this paper could be further strengthened if comparison was made between virtual and traditional urgent care providers.

On the other hand, we only studied the five most popular diagnostic groups, which made up 38.04% of total encounters, instead of analyzing all diagnoses. The reason for this is that the original dataset came with 124 types of diagnoses, which we were unable to categorized into meaningful groups without expert knowledge in disease classification. Therefore, further studies may focus on analyzing a selective group of diagnoses, which would yield more disease-specific findings.

Last but not least, the quality and reproducibility of telemedicine consultation remained in question. To our knowledge, only one study evaluated the reproducibility of telemedicine diagnosis of common acute problem, suggesting a strong trend for telemedicine physicians to disagree on primary diagnosis made by office-based physicians. Future work could test the inter-physician agreement on prescription outcome of a certain type of acute illness. Additional test could be designed to assess the unbiasedness of telemedicine provider, measured by the consistency of prescription outcome generated by an individual provider.

**Future Direction**

Previous studies highlighted the pivotal role of an extensive examination of both the quantitative and the qualitative determinants in order to understand the motivation and preference behind the physician’s prescribing behavior. Therefore, future research could incorporate qualitative metrics that measure personal, organizational, and sociotechnical qualities: the physicians’ attitudes toward telemedicine, work ethics, and organizational culture, etc. Such measures could be obtained through survey questions or contextual interview.

On the other hand, traditional urgent care and ambulatory care had been known for antibiotics overuse. According to one study, about 1 in 3 antibiotics prescribed at ambulatory care visits was unnecessary. Researchers could analyze the antibiotics prescription pattern in the telemedicine interventions, which will add to the current knowledge of antibiotic abuse.

Moreover, it was unknown whether the virtual care providers had received telemedicine-specific training prior to practicing online. We highly recommend telemedicine developers and healthcare organization leadership to consider developing training and educational programs in order to standardize and regulate prescription appropriateness in the virtual setting.
Conclusion

In conclusion, this study investigated physicians prescribing behaviors in a telemedicine urgent care platform. We identified that male primary care physicians made up the majority of the telemedicine provider population; this subgroup was also the most likely to prescribe a greater number of drugs at a given consultation. Primary Care providers and EM physicians showed different prescribing styles, with the latter appeared to be more reluctant toward prescribing online. However, this study alone could not explain the inter-physician variation in prescribing behavior due to the lack of access to qualitative measures such as the physician’s attitude toward practicing online.

Nevertheless, this study highlighted the difference in prescribing pattern among telemedicine providers, which could potentially compromise the quality of virtual care delivery. We urge future researchers to focus on the appropriateness and reproducibility of telemedicine physician’s prescribing outcomes.

While telemedicine has been increasingly perceived as a timely, cost-effective alternative to traditional office-based medical care delivery, it is crucial to ensure that the quality of care is not compromised. We believe that an enhanced understanding of telemedicine providers’ behaviors would help improving the delivery of around-the-clock, appropriate, and accountable medical care to the population.

References


Improving the Odds of Success for Precision Medicine Using the Social Ecological Model.

Scott P. McGrath, PhD
University of Nebraska at Omaha, Omaha, NE

Abstract
The concept of precision medicine aims to provide additional context to patient data for healthcare providers. A decade after the HITECH act of 2009, the state of EHRs can be considered a mixed bag. Increased levels of physician burnout have been attributed to the impact EHRs have had on traditional patient and provider interactions. In order for precision medicine to be allowed to establish a foothold, it must demonstrate the ability to improve clinical outcomes. One path to achieving this is by improving health behavior, which is a difficult task. In this paper, the case is presented for using the social ecological model to help shift health behaviors with precision medicine.

Introduction
In this era of start-ups, over-night success stories, and Internet billionaires, there is a tendency to oversell and overhype ideas and concepts; a condition the healthcare industry is not immune to. Take for example the company Theranos, a Silicon Valley health technology company which promised to revolutionize blood testing with its microfluidics, and earned its founder, Elizabeth Holmes, the title of the youngest female self-made billionaire. The company had been valued at $9 billion, but in October of 2015, an expose in the Wall Street Journal indicated the company wasn’t using its technology on the blood tests it was conducting for customers. A year later the company was embroiled in a lawsuit for $140 million from Walgreens. The FDA has banned Ms. Holmes from the blood testing business for two years, and as of August 31st 2018, Theranos has ceased operation. This provides a clear warning to those who are looking for miracle “silver bullets” or chasing after the next “killer app” in the medical domain: expediency may work for developing social media platforms and ride sharing apps, however efforts to find a one-size fits all solution for healthcare is often short sighted.

The concept of precision medicine (Fig.1) aims to provide additional context to patient data for healthcare providers. Precision medicine overlays three additional layers of patient data on top of standard patient information: environmental exposure, personal lifestyle and behavior patterns, and information gleaned from their genome. While precision medicine promises to provide physicians and their patients with increased levels of customization for preventative and primary care, there is evidence that preparation relating to the genomic pillar of precision medicine is not yet adequate to ensure successful deployment.

Genomics is not a panacea for all our ills, so to focus on how the genomic aspect of precision medicine can advance our knowledge and ability to prevent and care for disease, it needs to be viewed in the proper context. Studies have clearly demonstrated that simple awareness of increased risk of disease due to genetics has no significant impact on improving health behavior (1-3). So to herald the genetic component of precision medicine as a ‘miracle elixir’ may be disingenuous at best. We often discuss “curing” cancer, but this frames the question in an improper mindset, as many people equate cure as returning to a state where the individual is as good as new. It is possible that offering a “cure” removes the individual’s overall share as a stakeholder in his or her own health. If a patient falls ill, they can get treated, and they return to a normal state. Patient investment in this scenario is more indirect through cost of time and money to the patient. Contrary to this, improving a patient’s health behavior increases their percentage of ownership in their health. Diet and exercise require far more investment in time and energy than adding a few pills to their morning routine. In the long term, it has been postulated that that adjusting health behavior will pay more dividends than simply being the recipient of a treatment or “cure” for a disease. It is simply stated, but it has been quite the enigma for physicians, how to engage patients to shift this perspective. I think it is imperative to frame precision medicine as a method to ultimately improve health behaviors instead as path to delivering new treatments and improving outcomes. To be sure, both are beneficial aspects of precision medicine, however there is a unique opportunity to transfer emphasis of care to preventative over reactive care through precision medicine.
Figure 1: Pillars of precision medicine: Traditional medical records will be supported by supplementing medical data with information that can be curated into the following categories: 1) environmental 2) lifestyle, and 3) genomic. The environmental pillar will be collected based on environmental exposure an individual encounters in both work and life. Lifestyle pillar will be comprised of data from personal activity, exercise, and diet. The genomic pillar will be built using an individual’s genetic and genomic data. In combination, all three will help build a fuller picture of an individual’s health in the precision medicine model.

To support this perspective, I propose we aim to implement the model of precision medicine using the social ecological model and social determinants of health. The advantage of using an ecological model is that they contain a heuristic value by forcing the researcher to consider multiple levels of influence on the subject of their study. Rimer and Viswanath (4) outlined five principles for ecological models rooted in studying health behavior (Figure 1):

1. There are multiple levels of influence on health behavior
2. Environmental contexts are significant determinants of health behaviors
3. Influences on behaviors interact across levels
4. Ecological models should be behavior specific
5. Multilevel interventions should be most effective in changing behaviors

Figure 2: The social ecological model and social determinants of health. (5) Reprinted with permission from the National Academies Press, Copyright 2003, National Academy of Sciences.
Social Ecological Model

Rimer and Viswanath point to the advances made in tobacco control as evidence of success with ecological models in changing health behavior. The Surgeon General began to highlight the dangers of cigarettes back in 1964, which initiated the process of encouraging Americans to quit smoking (6). Economic factors, mass media education efforts, clinical smoking cessation programs, and multi-level programs have had an aggregative effect on lowering smoking prevalence (7). This is a good example of viewing the social ecological model working at the public health level. When addressing such a large population, 42.4% of all adults in America were smoking in 1965 (8), smoking cessation approaches needed to be launched on multiple layers identified in the social ecological model.

An additional example of the application of social ecological model to help address the health and behavior associated with a disease can also be seen in the control efforts to combat the spread of HIV infection. Efforts to address HIV-spread have been most successful when multi-level approaches were taken to address societal and structural HIV behavior change, in addition to preventing the further spread of virus (9-11). Effective interventions were identified as methods that “contained attitudinal arguments, educational information, behavioral skills arguments, and behavioral skills training” and the least effective methods centered on using fear. Target audience gender, ethnicity, risk levels, and other variables influenced strategy success (11). The complexity of the number of variables that can be leveraged to help shift health behavior can be difficult to track, which is where the socio-ecological model can help. Kaufman et al. used the ecological model to help categorize some of these variables.

Alignment of Precision Medicine and the Social Ecological Model

The proposed model of precision medicine highlights lifestyle, environmental, and genetic data. There are two primary influencers in the ecological model: external factors and internal factors. When you look at environmental and lifestyle data, these largely align to the outer layers of the ecological model. Environmental impacts are associated with where an individual lives, works, and travels. In order to impact or change the exposures encountered in various environments, it is often too large a task for just a single individual, therefore it is better suited for large organizations or governmental bureaucracies to engage as well. Later in this paper we will examine how the Learning Health System may be an effective tool for these outer layers of the ecological model. Individuals have a lot more influence on their lifestyle, based off what they elect to eat and their activity levels. However, just like smoking, social and economic pressures can help influence health behaviors. Implementing soda or sugar taxes have shown to be an effective method of helping people improve their dietary consumption patterns associated with chronic disease and obesity (12).

Genetic data fits squarely into the intrapersonal layer. Intrapersonal is an internal influence on behavior, which can impact the perceived environment. By keeping the focus on improving health behavior, the intention is to help keep resisting the impulse to have precision medicine emphasize treatment over prevention. The ecological model should help with this challenge by keeping the environmental and lifestyle framed in the picture. Research has shown that simple notification of genetic risk of disease has no significant effect on smoking cessation, diet, physical activity, or even preventative actions like alcohol use, or applying sunscreen (13). Meaning, simple disclosure of the risks does not lead to changes in health behaviors.

If coupled with other layers of the ecological model, odds of behavior change will likely be improved. These changes were demonstrated in a recent study inspecting the impact of biomarker, genetic testing, and lifestyle factors on health behavior for individuals who were first-degree relatives of patients suffering from rheumatoid arthritis (RA), but who were not suffering from RA themselves (14). Researchers found that participants who were provided with access to a web based Personalized Risk Estimator for Rheumatoid Arthritis (PRE-RA), which included discussion over genotype, autoantibody results, and behavior changes that could improve RA-risk (smoking, diet, exercise, or dental hygiene) showed positive behavioral changes. The control group for this study was provided with standard RA education. The intervention helped to improve health behaviors at six months with 63.9% of the PRE-RA group following the recommended changes in comparison to the 50% of the control group who had not received the non-personalized standard RA education. Taking guidance from studies like this can help keep our focus on developing precision medicine in a way that will help to improve clinical outcomes.
The social ecological model's emphasize that health behavior is not simply influenced in a binary fashion is what makes it an appealing approach to consider when trying to deploy precision medicine. Take the simple action of prescribing medication to treat a disease. The action is fairly linear in design, a diagnosis is reached, and a treatment plan is developed with the inclusion of a selected drug. However, a variety of elements contribute to the ultimate efficacy of the drug. Medication regimen adherence is one a major hurdle care providers are faced with, and adherence is influenced by multiple factors itself: cost of the drugs, how “forgetful” the patient is, side effects from the drug, etc. This is why precision medicine needs to be developed in a way that was considered as the diversity of influenced on health behavior and to also accommodate people’s lifestyles as well. Not all layers of the ecological model have the same level of weight that influences the individual. Social, family and community networks will help shape individual behavior more than broad national campaigns occurring at outer layer of the ecological model. As demonstrated with the campaign against smoking, approaching the problem from several different levels (interpersonal, social, living and working environments, and societal) put up clear boundaries that collectively helped changed health behaviors. We cannot claim precision medicine as a successful model if “wins” are only described in individual cases. The financial costs and work investments required to launch and sustain precision medicine requires sustained shifts in clinical outcomes at the population level to help justify its expense.

**Individual behavior impact**

In order to explore this framework as a model to help implement precision medicine, I have conducted a few studies working within the living and working layer of the ecological model as part of my dissertation. The focus of my dissertation was examining how to help a growing patient population understand genetic results as we see more genomic data being introduced to health records. The first project explored the knowledge gap between primary care physicians and genetic experts (clinical geneticist and genetic counselors) via a validated survey instrument. A national survey was conducted, and participants (n=264) were sorted into two categories, genetic specialists (genetic counselors and clinical geneticists) and medical providers (primary care, internists, physicians’ assistants, advanced nurse practitioners, etc.). The survey presented three genetic test report scenarios for interpretation: a genetic risk for diabetes, genomic sequencing for symptoms report implicating a potential HMN7B: distal hereditary motor neuropathy VIIIB diagnosis, and a statin-induced myopathy risk. Participants were also asked about their opinions on Direct-to-Consumer-Genetic Test results and rank their own perceived level of preparedness to review genetic test results with patients. The rates of correctly interpreting results were relatively high (74.4% for the providers compared to the specialist’s 83.4%) and age, prior genetic test consultation experience, and level of trust assigned to the reports were associated with higher correct interpretation rates. The self-selected efficacy and the level of preparedness to consult on a patient’s genetic results were higher for the specialists than the provider group. This study lends support that the genetic specialist still remains the best candidates to educate patients about genetic test results, however primary care healthcare providers can provide accurate interpretations when specialists are not available. Initial findings were shared at the 2019 AMIA Informatics Summit in San Francisco as a podium abstract, and a full manuscript is currently under editorial review for publication.

The second project was a scoping review to map out the ideal educational techniques for physicians to help increase patient comprehension of genetic testing results. Twenty-six manuscripts were selected for inclusion in the scoping. The papers were categorized into three themes that emerged: papers focused on education interventions between patients and providers, professional and curriculum development to enhance adoption and comprehension of medical genetics, and review papers, which were found to have overlap with the intended aim of this paper. There are a few clear areas where emphasis should be placed for further work on this topic: continuing to develop web-based educational tools designed with active learning methodologies, conducting more randomized controlled trials to develop more robust evidence, and linking clinical outcomes with educational interventions, which will in turn assist in developing new medical curriculum with a renewed emphasis on genomics. The full manuscript for this study is also under editorial review for publication with an available preprint publication online (15).

The final project documented the process of developing a mHealth app designed to aid health care professionals (certified athletic trainers) in detecting and managing concussions in patients under their care (16). Telehealth and mHealth are attractive technologies to help extend the reach of health professionals with genetic expertise. By exploring these various components, the goals of proposed my dissertation research was (a) to identify barriers facing healthcare providers in regards to educating patients about their own genes/genomes (b), to outline the best pedagogical approaches for patient education, and (c) highlight how existing health technology can assist with said communication. These three studies focused on how to assist healthcare providers in their efforts to increase patient
understanding when dealing with data generated from the genomic pillar of precision medicine. Additional efforts will be needed at different layers of the ecological model in order to help shape health behaviors. In the following section, and the learning health system will be explored as example of how to impact the social economic and cultural level of the model. Precision medicine has moved from theoretical and planning stages into real world initiatives. There are a large amount of resources already being invested to usher it into practice. It is critical to identify future roadblocks while they still remain manageable and investigate solutions; this work aspires to target a few of these roadblocks and present some countermeasures to help facilitate the deployment of precision medicine.

**Learning Health System**

Addressing ways to work within the broad social economic and cultural level of the social ecological model is challenging, however the Learning Health System is one potential candidate to work in this space. One of the biggest impediments to advancing medical care in the United States is the extremely protracted length of time it can take to innovate medical procedures and to align the best medical practices with the most current corpus of clinical knowledge. It takes, on average, 17 years for research findings to be translated into products, policy, or clinical practice (17). How can we accelerate this pipeline to improve clinical care in the near future, and not in the next decade or two, what strides have been made to reach this goal? The answer is we have not significantly moved the needle on the “time lag” for biomedical research into clinical practice. The period of time remains between 10 to 25 year depending on which medical research domain is being evaluated (18), which is still too long. One proposed solution is the learning health system (19-21). Under the Learning Health System (LHS) model, a federated system of health data derived from EHRs would be established at the national level (19). This would allow members of the LHS to pose queries, to which others can supply relevant responses from their own collected data. This could occur in a variety of ways. For example, in drug trials, participants could be recruited by asking within the LHS “How many patients who meet these specific eligibility criteria does your institution have?” subsequently, if the drug reaches the market, the LHS can help monitor the drug’s safety by cataloguing and adverse health events. Real-time national surveillance can be enhanced through tracking of infectious disease. In fact, the entire aim of a LHS is to deliver “a health system that learns as quickly as possible about the best treatment options appropriate for each patient—and delivers it.” (22). The implications of the LHS is also recognized by the Health and Medicine Division, formerly the Institute of Medicine (IOM), which state “the nation’s health and economic futures—best care at lower cost—depend on the ability to steward the evolution of a continuously learning health care system” (23). This belief was also adopted by the Office of the National Coordinator for Health Information Technology (ONC) as they set it as a part of their 10-year strategic goal where a LHS would help to reduce health care costs, improve public health, empower patients, and help usher in new innovation (24). The need for this is urgent, with as much as 25% of medical expenditures, an approximate $750 billion per year, is lost due to administrative inefficiencies, abuse, fraud, and treatments which failed to improve clinical outcomes (20). Even with an expenditure of 18% of the US gross domestic product on healthcare (25) we still see patients who fail to receive 45% of recommended care, and overall healthcare quality improvements remains anemic (20). LHS is being recognized as a major plank in reducing the ‘GDP footprint’ healthcare currently occupies (26).

LHS is strong candidate to help drive precision medicine forward, it can also serve as an element that helps to extend beyond the individual level of the social ecological model. This partnership has been recognize by others, as outlined by Dr. Chambers in his 2016 article (27). Precision medicine is positioned to generate a sizable amount of data to help with clinical decision-making, but ultimately the means to measure the level of clinical effectiveness of the data in a cost efficient manner is lacking. This is where the LHS can assist, by refocusing on organization’s goals to ongoing system improvement in an incremental process, much like the agile software development model. The pairing of LHS and precision medicine is a natural fit, but the addition of implementation science is needed to bridge the two components. Implementation science is defined as “the scientific study of methods to promote the systematic uptake of research findings and other evidence-based practices into routine practice” (28). Discoveries and breakthroughs identified by precision medicine can be plugged into the LHS via a framework provided by implementation science (Fig. 2).
Health care services are housed within the living and working strata of the ecological model (5), a LHS could serve as a conduit for precision medicine have an impact at that layer. A nationalized LHS could also eventually reach the outer layer of the model by helping advance public health with enhanced surveillance and improved clinical treatment plans and outcomes.

The LHS concept was initially proposed back in 2007 (29), so how is the model coming along? The answer is we have seen several organization making progress toward a LHS (Geisinger (30), Kaiser Permanente (31), and the VA (32)) and we are reaching a point where impact can start to be measurable/quantitated, but there appears to be a lack of research on evaluating the impact (33). Efforts appear to still be largely in arguing the feasibility of the LHS. In a recent publication, Embi et al presented some policy recommendations to help LHS move past this argument by establishing the “evidence-generating medicine” phase (EGM) (34). As defined, EGM is the “systematic incorporation of research and quality improvement considerations into the organization and practice of healthcare in order to advance biomedical science and thereby improve the health of individuals and populations.” (35). Three principal recommendations were put forth by Embi: evidence needs to be generated at the local level (point-of-care), research needs to be conducted across organizations, and this ecosystem of local data generation and nationalized research needs to be supported and allowed to grow. To accomplish increased data generation, research should be incentivized at the local level, and regulations adapted to foster this aim, which require additional investment in health IT. Networking research between organizations requires a skilled workforce that is trained in informatics, and standards need to be set in addition to establishing governance protocols. Sustaining progress achieved through the first two aims will require strategic adjustments of federal funding guidelines and there should be an effort to improve the use of the research data, allowing for multiple parties to extract value.

Conclusions

The social ecological model helps us to understand that health behavior can be impacted by multiple factors, and is often not a single variable that leads people to make the decisions they make. It is possible to influence health behavior in part by focused efforts that only engage the individual at a single layer of the ecological model, but evidence points to efforts that try to engage individuals on multiple layers are more effective. Simple notification of genetic risks does not lead to lasting change, which is why precision medicine really needs to be viewed as an enhancement to current medical care instead of a re-inventing of care. The lifestyle, environmental, and genetic data accumulated into patient records should be treated like a spice or seasoning that is present to enhance the quality of the meal, but not to replace or overwhelm the meal itself. Treatment and patient engagement with the goal of
changing health behavior can be enhanced with blending in some of this information to find the correct balance for subgroups of patients. Some of the data we generate through precision medicine will have universal application, much like salt and pepper, others will have more nuanced applications where it will only need to be used sparingly, like saffron, in order to make the desired impact. Therefore, genetic risk assessment scores are expected to develop into the type of ‘seasoning’ that will have limited use cases. It isn’t something that you want to roll out for every patient, but when properly identified, it can lead to very positive results. Pharmacogenomics and expanding our treatment options will be more like the general ‘seasoning’, where large number of individuals will benefit from increased choices. The social ecological model helps us to frame precision medicine in the scope and scale suitable for our effort to improve health behaviors. Efforts require attention to the details, knowing when to pursue what proves to actually be effective, and avoid chasing after silver bullets or overhyped claims.

References

15. McGrath S, Bastola K. Educating patients on genomic results: a scoping review. JMIR Preprints. 22/02/2019:13799. DOI: 10.2196/preprints.13799. URL: https://preprints.jmir.org/preprint/13799
A Qualitative Study to Examine Approaches used to Manage Data about Health Facilities and their Challenges: A Case of Uganda

Jonathan Mpango, MS¹, Josephine Nabukenya, PhD²
¹School of Public Health, Makerere University, Kampala, Uganda; ²School of Computing and Informatics Technology, Makerere University, Kampala, Uganda

Abstract

Availability of an accurate and complete health facility list is fundamental in producing quality and timely data that is sufficient to aid evidence-based decision, resource allocation and planning within the healthcare ecosystem. This study aimed at examining the approaches used in Uganda to manage data about health facilities and the challenges they are facing. We conducted a qualitative study involving 32 interviews with participants from Ministry of Health, government regulatory organizations, district local government, general public, academia, implementing partners and healthcare providers. Our analysis identified four divergent approaches that had five common challenges, namely: lack of a health facility unique identifier, non-standardized, incomplete, inaccurate data, difficulty accessing and using data. Establishing a national central health facility registry to manage the national health facility list would improve patient referrals, facility look-ups, health information exchange, data curation and access and health information system integration.

Keywords: Health Facilities, Health Information Exchange, Master Facility List, eHealth, Uganda

Introduction

Availability of accurate and timely information on the quality and supply of healthcare services is essential for monitoring and evaluating health systems (1). Having quality and complete information enables countries to track the progress and performance of their health information systems (HIS). However, few countries have accurate and up-to-date information on the state of their health facilities (HFs), covering the public, private-for-profit and private-not-for-profit sectors. This poses a challenge in producing quality and timely data that is sufficient to aid evidence-based decisions and resource allocation (2). One innovative approach that low and middle-income countries (LMICs) have undertaken is embracing Information and Communication Technology (ICT) to develop and implement electronic HIS (eHIS) within the healthcare system to manage data about HFs (3).

Despite the huge investment in implementing eHIS in Uganda, it is still challenged by divergent levels of data standards that hinder data exchange, poor coordination and communication, limited skilled human resource, fragmented data collection processes, inadequate infrastructure and isolated HIS (4–6). According to World Health Organization (WHO), most of the LMICs have multiple sets of HF lists which may contain contradicting data on HFs and they lack a common HF unique identification number (7). Maintaining such multiple HF lists is costly and in case they contradict each other, it may be difficult to identify the correct one. To address this challenge, WHO recommends that countries should establish one comprehensive Master Facility List (MFL) that should house complete, accurate and up-to-date data on all HFs that exist in a given geographical location (7,8). This recommendation is also mentioned in the Uganda eHealth Policy (9).

In recent years, progress has been made globally by different countries to develop tools that support them to collect and manage data about HFs (7,8,10–12). Several studies have mentioned the benefits that have been attained by countries that have established a comprehensive and up-to-date MFL including; supporting emergency response operations, efficient allocation and distribution of resources related to HFs, improved disease surveillance and improved HF identification (7,8,10–12). The government of Uganda through its in-line Ministry of Health (MoH) in collaboration with its stakeholders has made progress in establishing a national MFL, which was yet to be officially launched by the time this study was conducted. Over time, the status of HFs in a given location change due to establishment of new ones, others get upgraded or downgraded whilst some are closed. This requires the MFL to be continuously updated whenever such changes occur, which are not an easy task.

In order to address this challenge, WHO recommends that countries can establish a national Health Facility Registry (HFR) to manage its MFL (8). The HFR is a software application that acts as the central authority to collect, store and distribute accurate, complete and standardized data on HFs (7,8,11). There are numerous benefits that are attained by countries that have established a national HFR that are documented in the existing literature. Implementing an HFR fosters better monitoring and evaluation by supporting the harmonization of facility linked datasets, improves referrals and facility look-ups, facilitates health information exchange (HIE), and facilitates data curation (7,13–15). Notwithstanding such benefits, there is no national HFR yet developed to manage Uganda’s MFL although some steps have been taken through the national eHealth strategy.
At the time this study was conducted, there was no evidence published in existing literature about the approaches used within the healthcare sector to manage and/or access data about HFs. Thus, we aimed to use the study findings to improve on the current approaches by establishing optimal requirements and features a HFR should have to better manage data about HFs in Uganda’s healthcare system.

Methods

Study settings: The study was conducted in Uganda, a landlocked country in the eastern region of Africa. Its estimated population is 42.86 million by the end of 2017, with an annual growth of three percent (16). Uganda healthcare system is decentralized composed of public and private sectors (17,18). The public sector is decentralized at district and sub-district levels ranging from national referral hospitals (NRHs), regional referral hospitals, general hospitals, health centers, to village health teams whilst the private is divided into private not for profit organizations (PHFPs) and private for profit organizations (PFPs).

Study design: A qualitative exploration approach was used in this study. An exploration approach is research approach used to investigate a problem that has not been studied more clearly which is the case with this study (19). The purpose was to examine the approaches that were being used to manage and/or access data about HFs within the Uganda healthcare sector.

Participant selection: Purposive sampling was used to identify the study participants from organizations that utilize data about HFs during their decision making process. This method of sampling helps to create a homogenous sample of participants that have experience with the phenomenon (20). Participants selected from the sampled organizations were required to at least have used data about HFs in their day-to-day operations. The category of participants selected included policy makers, monitoring and evaluation officers, technical officers, program officers, district health officers, district biostatisticians, academic researchers, healthcare providers, data officers and patients. In this study, participants were no longer recruited once saturation had been reached, in which no new information was provided to aid the understanding of the phenomenon (21). Participants were contacted face to face on appointment with a cover letter and consent form explaining their rights as participants.

Data collection: The first author, who completed data collection, bracketed biases before beginning data collection to assure data accuracy. A semi-structured interview protocol was developed and reviewed by both authors, and edited based on feedback. The interview questions were generated from the literature review around the phenomenon under study. These questions were reviewed and modified by consensus between the authors who have vast knowledge on HIS and the healthcare system of Uganda. Each interview was treated as an individual case, and all interviews were audio-recorded and de-identified. Throughout the process of the interviews, probes and follow-up questions were added as needed to encourage amplification and elucidate responses. Specific questions were added as the interview process proceeded. Data collection was completed over a period of 2 months and ended upon saturation.

Table 1. Semi-structured interview protocol

| What approach do you use to manage data about health facilities at your organization or national level? |
| Probe: What other approaches are you aware of in the sector?; Currently, how do you access data about health facilities in the approaches you have mentioned? |
| What issues affect the mentioned approaches for managing data about health facilities? |
| What are your recommendations to address the mentioned challenges? |

Data analysis: At the end of the interview process, the audio recordings were transcribed verbatim by the first author and validated independently by the co-authors who re-listened to the audio recordings to ensure they were accurately transcribed. Observations from the co-authors were shared with the first author who incorporated them into the final interview script. We engaged in a reflexive dialogue to conduct an open inductive analysis (22,23). After familiarization with the raw data a coding scheme was developed in a multi-level process. Statements that mentioned approaches, challenges and recommendations were identified and categorized into codes. Any discrepancies were discussed and new codes or code definitions were created. Trustworthiness of findings was enhanced through frequent discussion between the authors to ensure that the codes, sub-themes and themes adequately described and encompassed data collected.
Study period: Data collection was done in August and September 2018 and analysis was done in October and November 2018.

Ethical considerations: The Higher Degrees, Research and Ethical Committee (HDREC) of Makerere University School of Public Health approved this study. All participants gave written consent and their privacy and confidentiality was ensured during the study.

Results
A total of 32 individuals involved in using data on health facilities consented to participate in our qualitative study, where 32 one-on-one key informant interviews were conducted with participants. Participants included 2 policy makers, 5 monitoring and evaluation officers, 7 technical officers, 3 program officers, 2 district health officers, 3 district biostatisticians, 1 academic researcher, 2 healthcare providers, 2 data officers, and 5 patients (Table 2).

Table 2. Study participants

<table>
<thead>
<tr>
<th>Organization</th>
<th>Participant category</th>
<th>Number of participants per category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ministry of Health</td>
<td>Policy makers</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Monitoring and evaluation officers</td>
<td>3</td>
</tr>
<tr>
<td>Government Regulatory Organizations</td>
<td>Technical officer</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Program officer</td>
<td>1</td>
</tr>
<tr>
<td>District Local Government</td>
<td>District Health Officers</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>District biostatisticians</td>
<td>3</td>
</tr>
<tr>
<td>Academia</td>
<td>Academic researcher</td>
<td>1</td>
</tr>
<tr>
<td>Development Partner</td>
<td>Technical officers</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Program officers</td>
<td>2</td>
</tr>
<tr>
<td>Implementing Partner</td>
<td>Monitoring and evaluation officers</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Technical officers</td>
<td>3</td>
</tr>
<tr>
<td>Health Facility</td>
<td>Healthcare providers</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Data officer</td>
<td>2</td>
</tr>
<tr>
<td>General Public</td>
<td>Patients</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>32</td>
</tr>
</tbody>
</table>

The study examined the approaches used to manage and/or access data on HFs and their challenges in Uganda. We present our findings grouped in two major themes that were identified during data analysis. These include: (i) approaches used to manage and/or access data about HFs, and (ii) the challenges faced by the identified approaches.

Theme 1: Approaches used to manage and/or access data about health facilities

Based on the data analysis, four approaches for managing and/or accessing data about HFs in Uganda healthcare sector were identified, namely; the draft 2018 Ministry of Health Master facility list, the national electronic health information system, the institutional-based facility list and the paper-based approach.

The 2018 Draft Ministry of Health Master Facility List: About 9% (3/32) of the study participants alluded to the 2018 draft MoH MFL as an approach used to manage data about HFs in the country. The 2018 draft MoH MFL is a Microsoft excel-based file that houses data about HFs (public and private) that has been compiled and managed by MoH in collaboration with its stakeholders. By the time this study was conducted, the MFL was yet to be released for public use. Some of the study respondents view it as the primary source of truth for data on HFs in the healthcare sector. The data attributes captured about a HF in this approach include, HF name, level, type, administrative location, ownership, operating status, contact address, geographic coordinates and services offered.

One of the comment from a study participant:
“The MoH has just compiled a national MFL that contains information on all health facilities in the country. The list has been validated and I look at it as the primary source of accurate and complete data on health facilities in our healthcare sector.” -- Participant 07, Monitoring and evaluation officer

The National Electronic Health Information System: The national eHis is a web-based application based on the open source District Health Information System (DHIS2) which is used to collect, store, and disseminate aggregated healthcare data to support data-driven decision making (4). HF's that are expected to report on the aggregated healthcare data are registered into the eHis. Thus, the eHis was mentioned as an approach used to manage and/or access data about HF's within the healthcare sector by 53% (17/32) of the study participants. The data attributes captured about a HF in the eHis include: HF name, level, ownership and geographic coordinates.

Some of the comments from the study participants:

“All health facilities are expected to report data into the national eHis, we consider the health facility list in the eHis when planning for activities that involve health facilities ...” -- Participant 04, Program officer

“... we often use the facility list we downloaded from the national DHIS2 when seeking for information about any specific health facility. We trust the information in the system because it is managed by the Ministry of Health...” -- Participant 08, Technical officer

The Institutional-based Facility List: The results show that 12% (4/32) of the study participants referenced to their own institutional-based facility lists for data about HF's. Several non-government and faith-based organization exist in Uganda healthcare sector and these support the MoH in strengthening healthcare service delivery especially at HF level. These organizations have established and maintained their own HF lists that they reference to during planning for activities that involve HF's. These institutional-based facility lists are managed using Microsoft office applications and access to them requires authorization from the managing institution. Majority of the institutional-based facility list observed during this study had varying data attributes collected to define a HF.

Some of the comments from the study participants:

“We have our own facility list that we routinely update with data on health facilities based on information we get when we physically visit these health facilities for on-site supervision and training activities.” -- Participant 02, Monitoring and evaluation officer

“... at the district, we have our own health facility list with all health facilities in our district that we manage.” – Participant 10, District biostatistician

The paper-based approach: Some of the study respondents especially the patients in the general public mentioned referencing to data about a given HF that they manually wrote on a piece of paper. The piece of paper was alluded to by 6% (2/32) of the study participants as an approach they use to manage and access data about HF's within the healthcare sector.

To quote a representation of these sentiments:

“I don’t have a system I refer to for information about a health facility. After getting a recommendation from a friend who received a service from a particular facility, I always write down the facility name, address and doctor’s name on a piece of paper which I refer to later before seeking healthcare.” -- Participant 13, Patient

Theme 2: Challenges faced by the identified approaches

Based on the results obtained, each of the identified approach used to manage and/or access data about HF's in the country had notable challenges that were mentioned by the study participants. These included: (i) lack of standardized HF unique identification number, (ii) non-standardized data attributes, (iii) incomplete HF listing, (iv) inaccurate data about HF's, (v) lack of policies and legislation that govern the facility listing, (vi) difficulty in accessing and using data about HF's.

Lack of standardized health facility identification number: 94% (30/32) of the study participants were of the general view that there was no standardized and approved HF unique identification number (UIN) that uniquely identified a HF from the rest within the healthcare sector. The 2018 MoH MFL had two UINs that were assigned to a HF (Table 3). The MoH did not yet approve both UINs by the time this study was conducted. 15% (17/32) of the study participants who mentioned the HF listing in the national eHis were not sure if it had a HF UIN. Although one of the study respondent mentioned that the eHis automatically generated and assigned a 128-bit system based globally unique identification number (GUID) whenever a new HF record is added into the system. The GUID is neither recognized nationally as an approved HF UIN. None of the study participants that mentioned the institutional-based facility list and the paper-based acknowledged that they had a HF UIN.
Some of the comments from the study participants:

“... the National Health Provider Identification (NHPI) code is an 8-digit code that is permanently assigned to a health facility and the code didn’t have any meaning attached to the health facility whilst the Health Service Taxonomy Code (HSDTC) was generated and assigned to a health facility based on its geographic location ... when a health facility location changes, a new HSDTC is generated and reassigned to that health facility.” -- Participant 10, Technical officer

“... the task of generating an d updating the two codes was complex ... and the codes that were assigned to the health facilities were too long to be remembered and used. We cannot use them at the district ...” Participant 14, District Health Officer

“The facility list does not have a unique identification code for the health facilities” -- Participant 02, Monitoring and evaluation officer

“No, I don’t include a unique identifier for a health facility... the health facility name and level is all I need to identify a health facility” -- Participant 13, Patient

Table 3: The proposed structure of the health facility unique identification number in the 2018 draft MoH MFL

<table>
<thead>
<tr>
<th>Administrative Location</th>
<th>NHPI Format</th>
<th>HSDTC Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>National (A)</td>
<td>UG</td>
<td>UG-256</td>
</tr>
<tr>
<td>Regional (B)</td>
<td>RG-XXXXC</td>
<td>A/NNN</td>
</tr>
<tr>
<td>District (C)</td>
<td>DT-XXXXC</td>
<td>A/B/NNN</td>
</tr>
<tr>
<td>County (D)</td>
<td>CN-XXXXC</td>
<td>B/C/NNN</td>
</tr>
<tr>
<td>Sub-county (E)</td>
<td>SC-XXXXC</td>
<td>C/D/NNN</td>
</tr>
<tr>
<td>Parish (F)</td>
<td>PA-XXXXC</td>
<td>D/E/NNN</td>
</tr>
<tr>
<td>Health Facility (G)</td>
<td>HF-XXXXC</td>
<td>E/F/NNN</td>
</tr>
</tbody>
</table>

X – 3-digit random alphanumeric code
N – 3-digit incremental numeric number
C – Numeric check code

Non-standardized HF data attributes: According to the data analyzed each approach had different set of data attributes that were collected to define a HF. The draft 2018 MoH MFL captured the HF name, unique identification number, level, type, ownership, administrative location (national, region, district, health sub district, county, sub county and parish), geographic coordinates, list of services offered, contact address, operational status and date when the data was collected as the HF data attributes. In the paper-based mechanism, patients mainly capture the facility name, physical address, contact of the healthcare doctor whilst the institutional-based facility lists had varying data attributes. The eHIS captured the HF name, level, ownership and geographic coordinates.

Some of the comments from the study participants:

“The MFL has data on health facility UIN, facility name, level, ownership, type, location, geographic coordinates, services offered, contact address, operation status and record date.” -- Participant 07, Monitoring and evaluation officer

“The list contains the facility name, level, ownership, type, administrative location and in-charge contact address” -- Participant 10, District biostatistician

“The national eHIS contains the health facility name, type, level, ownership and geographic coordinates as attributes collected on a health facility...” -- Participant 04, Program officer

Incomplete HF listing: The completeness of the HF listing in the identified approaches varied across the study participants. Some felt that the national eHIS lacked data on majority of the private HFs in the healthcare sector. The institutional-based facility list mostly contained information about HFs that exist in the locality where that particular institution/organization operate leaving out the rest of the HFs. According to some of the study participants, the 2018 draft MoH MFL had complete listing of all HFs (private and public) that existed in the
healthcare sector since the list underwent several validations with different stakeholders. The paper-based approach contained only data about HFs that are of concern to a particular individual/patient.

To quote a representation of these sentiments:

“the eHIS facility listing does not include all health facilities in the health sector; most of the private health facilities are not included since they are not obligated to report through the reporting system” -- Participant 17, Healthcare provider

Inaccurate data about HFs: Respondents reported that some of the attributes captured on a HF in the national eHIS were not accurate whilst those who use the institution-based facility list mentioned that they routinely update the data about HFs based on information they receive from the district health officer and the information they get when they physically visit the HFs during on-site supervision and training activities. The respondents didn’t not mention about the accuracy of data in the individual paper-based approach, however, majority of data about HFs is obtained through recommendations from a friend, workmates or relative.

For example, some of the study participants mentioned:

“...the geographic coordinates in the national eHIS for some of the health facilities are not accurate and we cannot rely on them during visualization of health facilities using the system GIS application” -- Participant 11, Monitoring and evaluation officer

“...after getting a recommendation from a friend...”-- Participant 13, Patient

Lack of policies and legislation that govern the facility listing: Majority of the respondents interviewed were very categorical on the need to establish some form of policy and legislation to guide on the process of managing and maintaining the data on HFs because they were missing.

To quote a representation of these sentiments:

“... district health team should be legally bound to maintain the national health facility list up-to-date with accurate information ...” -- Participant 14, District health officer

“...financing of districts by MoH should be based on the total number of health facilities in a particular district that are accurately maintained in the national health facility list. ...” -- Participant 05, Academic researcher

Difficulty in accessing and using data about HFs: With a few exceptions, the general consensus among the respondents was that accessing data about HFs across the different identified approaches was difficult. To access data about HFs in the eHIS and the institutional-based approaches require authorization and authentication from MoH and the institution managing the facility list and MoH respectively. The 2018 draft MoH MFL was yet to be launched, therefore the data was not yet available for public use at the time this study was conducted.

To quote a representation of these sentiments:

“I lost my cousin because we were in a location that we weren’t familiar with and we couldn’t tell which health facility was nearby with a particular specialization to save life. If the country had a comprehensive health facility list with all details on facility location, facility contact and services offered at a particular health facility that we can freely access, then a life could have been saved.” -- Participant 05, Academic researcher

“...before implementing any activity regarding health facilities in a specific district, we first inquire from the district health team who then share with us their district specific health facility list stored in Microsoft excel ...” -- Participant 09, Monitoring and evaluation officer

Discussion

This study provided an insight into the current approaches used in Uganda to manage and/or access data about HFs and the challenges they are facing from the perspective of the study participants. The key findings revealed that multiple approaches for managing and/or accessing data about HFs existed and they had notable challenges. A country having multiple approaches for managing data about HF is not new in existing literature (12,24); The approaches established by this study included the 2018 draft MoH MFL, the national eHIS, institutional-based facility list and the paper-based mechanism. The national eHealth policy mentions of a need to merge all data from all the existing approaches and establish one national MFL that is centrally managed by MoH and accessed by all stakeholders in Uganda healthcare sector (9). This recommendation is in line with the desired recommendation given by the study participants.

This study also revealed that all the identified approaches had notable challenges. These included, lack of standardized HF unique identification number, non-standardized HF attributes, incomplete HF listing, inaccurate data about HFs, lack of policies and legislation that govern the facility listing, and difficulty in accessing and
using data about HFs. In the next sections of the paper, we discuss the challenges identified with the approaches currently used to manage and/or access data about HFs. We further present the electronic Health Facility Registry (HFR) for managing the MFL as a potential intervention for addressing the challenges observed and improving the efficiency of the HF listing efforts in the country.

All the identified approaches lacked a standardized and approved UIN that uniquely identified a HF in the healthcare sector. This is attributed to the fact that the central governing body (MoH) has not yet established a standard one to uniquely identify HFs within the healthcare sector. This has led to the establishment of an enterprise-wide system of non-standardized identification numbers that are generated and assigned to HFs by different entities. Establishing a common standard UIN for HFs has numerous benefits as observed in literature. These include: improved efficiency in HF identification, management and resource distribution (8,10,25); creates opportunities for HIS interoperability (25); improves healthcare referrals (26–28); and facilitates exchange of data on health facilities (29,30). Nigeria has implemented a HFR and the HF UIN has meaning associated with the current government area and state of location of the HF (25). When a HF moves to a new government area or state, the existing UIN is nullified and a new one is generated. As a result, the HF UN becomes burdensome to use for various stakeholders (doctors, district planners, program officers, among others) when identifying a HF. Our recommendation from the study participants is a HF UIN that is short, clear, precise and not associated its current geographical location.

The data attributes collected on each health facility record were different across the identified approaches. This hinders achieving data exchange and system integration (8,10,25). According to WHO guidelines for establishing a national MFL, a MFL should contain the HF unique identifier, facility name, facility type, ownership, administrative location, geographic coordinates, operational status, facility contact address, data collection, services offered, available human resource and available inpatient and maternity beds as the minimum dataset collected on a HF record (7,8). According to Rose-wood et al., having an accurate and complete MFL within a healthcare sector is fundamental for effective planning, coordination and delivery of health services (12). By the time this study was conducted, none of the identified approaches was nationally recognized to house accurate and complete data about HFs within the healthcare sector. The situation of having multiple HIS in LMICs and none serves as a national repository for accurate, up-to-date and complete listing of health facilities is not new in literature (8,12,31). There is need to establish a national MFL that has accurate, up-to-date, complete, and standardized data about HFs that exist within the healthcare sector.

Existing literature reveals that unavailability of standard policies and legislations to govern access and usage of data about HFs, deficiencies in skilled human resource to manage the data about HFs, and inadequate implementation of the recommended practices for data management are some of the reasons that can be attributed to the discrepancies in data about HFs (25,32). This study has revealed that there were no policies that defined how data about HFs should be established, updated, deleted and archived. To maintain the data about HFs accurate, up-to-date and complete, there is a need to establish and enforce proper legislation aimed at ensuring that data is regularly updated (8,25). Additionally, a data governance framework is needed to manage data issues such as security, sharing and access associated with the MFL (10,12).

Accessibility and availability of data about HFs housed in the MFL is a fundamental factor to improved health service delivery and effective distribution and planning of resources within a healthcare sector (4,33,34). The study finding revealed that majority of the approaches required authorization to access and use the data about HFs. This could have led to the establishment of multiple approaches for managing and/or accessing data about HFs in the sector. WHO recommends that the data about HFs should be easily accessible to all stakeholders at any given period of time to aid data use for data-driven decision making, monitoring of HF distribution and resource allocation, HF identification and patient referrals (8).

To address the challenge of maintaining and managing the MFL with accurate, complete and up-to-date data, a HFR has to be implemented (25,35,36). The HFR can be used to manage and/or access data about HFs in Uganda to aid evidence-based decision making at all strategic levels within the healthcare sector. With the HFR established, other information systems and/or stakeholders can easily access its data. Future research could focus on establishing requirements for establishing a functional HFR in Uganda.

**Conclusion**

Our study has provided evidence on existence of various approaches used for managing and accessing data on health facilities in Uganda. The study has also revealed some important challenges that hinder effective use of these approaches including lack of a health facility unique identification number, management of data on health facilities, health information exchange and interoperability because each mechanism had different data elements captured, and data access and use to aid evidence-based decision making. As such, we suggest that in order to improve the task of managing and maintaining data on health facilities accurate, complete and up-to-date, there is need to establish a national HFR as it was alluded to by the national eHealth policy (9). The government of Uganda
through its line Ministry of Health should improve on its in-house coordination and governance including establishment and implementation of policies that should focus on eliminating duplicated approaches established by various health institutions and promote the establishment of a central HFR that can be accessed and used by all stakeholders in the healthcare system.

Acknowledgment
The authors would like to thank the respondents who generously extended their time for the interview and openly shared their opinions on the subject under study. Without them it would have been impossible to accomplish this work. In addition, we acknowledge the moral support from Makerere University School of Public Health and Monitoring and Evaluation Technical Support (METS) Project and the financial support from the HI-TRAIN program, a Norwegian funded project for higher education in Uganda for conducting this research.

Statement on conflicts of interest
There is no conflict of interest that the authors need to declare.

References


progress-on-a-master-facility-list-and-health-facility-registry/


Process Mining and Ethnography Study of Medication Reconciliation Tasks

Vaishak Ramesh Vellore, MCS¹, M. Adela Grando, PhD², Benjamin Duncan, MS², David R. Kaufman, PhD², Stephanie K. Furniss, PhD²,³, Bradley N. Doebbeling, MD, MSE²,⁴, Karl A. Poterack, MD³,⁵, Timothy Miksch, MBA, MAS³, Richard A. Helmers, MD⁶

¹ Fulton Schools of Engineering, Arizona State University, AZ, US; ² Biomedical Informatics, College of Health Solutions, Arizona State University, AZ, US; ³ Kern Center Informatics and Knowledge Management, Mayo Clinic, Rochester, MN, US; ⁴ Science of Healthcare Delivery, Arizona State University, AZ, US; ⁵ Department of Anesthesiology, Mayo Clinic, AZ, US; ⁶ College of Medicine, Mayo Clinic, Rochester, MN, US.

Abstract
We studied the medication reconciliation (MedRec) task through analysis of computer logs and ethnographic data. Time spent by healthcare providers performing MedRec was compared between two different EHR systems used at four different regional perioperative settings. Only one of the EHRs used at two settings generated computer logs that supported automatic discovery of the MedRec task. At those two settings, 53 providers generated 383 MedRec instances. Findings from the computer logs were validated with ethnographic data, leading to the identification and removal of 47 outliers. Without outliers, one of the settings had slightly smaller mean (SD) time in seconds 67.3 (40.2) compared with the other, 92.1 (25). The difference in time metrics was statistically significant (p<.001). Reusability of an existing task-based analytic method allowed for rapid study of EHR-based workflow and task.

Introduction
Medication errors are a leading cause of patient harm. In over half of the patients, discrepancies were found between the medication patients were taking at home and the list of medications known to the hospital caregivers after intake.(1–3) Many of these errors result from an incomplete overview of medication either at a patient’s referral to or at discharge from the hospital. Medication Reconciliation (MedRec) is the process of creating the most accurate and complete list of medications a patient is taking—including drug name, dosage, frequency, and route—and comparing that list against the physician’s admission, transfer, and/or discharge orders, with the goal of providing correct medications to the patient at all transition points within the hospital.(4)

MedRec plays an important role in reducing adverse drug events (ADEs). A common approach to estimate the impact and effectiveness of MedRec on ADEs has been to use database entries and EHR data, which were then evaluated independently by experienced physicians.(5,6) Several studies used observational data to assess the processes required to carry out MedRec and estimate severity of ADE’s in-order to develop ADE prevention strategies and ascertain those related to medication errors.(7,8) Interviews have also been used to assess providers’ perspectives on the use of technology for MedRec to improve ambulatory medication safety and identify barriers of the implementation process.(9,10)

Medication safety could be affected by the amount of time a professional spends on MedRec. Tamblyn et. al developed a prototype e-medical reconciliation software in the surgical wards to measure the time spent on MedRec when discharging patients per professional.(11) The
overall Mean (SD) time in minutes during each of the RightRx session was 5.8 (7.8) for physicians and 10.9 (10.9) for pharmacists. Meguerditchian et. al used a time-and-motion study to observe and time health professionals in three hospital units: geriatrics, medicine and surgery.(12) Only one major task, preparing the discharge prescription, was selected for timing during MedRec at discharge. The results showed that the difference in time spent on MedRec varied greatly between units, ranging from an average of 10 minutes (general surgery) to 92 minutes (geriatrics).

Mayo Clinic has undergone an enterprise-wide, large-scale EHR conversion. Before the conversion, the ROOT (Registry Of Operations and Tasks) project was launched in 2016 to document and harmonize EHR-mediated workflows and Health Information Technologies (HIT) interactions. ROOT focusses on understanding variations across different systems and sites by using broad range of methods including video ethnography, computer logs analysis and interviews.

ROOT’s approach consists on triangulating ethnographic methods and automatic analysis (mining) of computer logs to understand different dimensions of EHR-mediated workflows.(13–16) This novel method was used to analyze and compare time metrics for the vital signs charting task before conversion to the new EHR. Vital signs charting was documented with two different EHRs used at four different preoperative (PreOp) regional clinic settings at the Mayo Clinic: Arizona (Phoenix, AZ), Jacksonville (Jacksonville, FL), Methodist and Saint Marys (Rochester, MN). Time metrics for vital sign charting were automatically quantified using computer log analysis. Outcomes were validated with ethnographic data and outliers could be identified, explained and removed. Without outliers, similar mean (SD) times in minutes (1.3 (1.1), 1.2 (1.3) and 1.2 (1.2)) were found between sites when performing vital sign charting.(17)

Here, we reused the same methodological approach applied to the study of vital sign charting to the analysis of a different task, MedRec. This study also occurred at the same settings where the vital sign charting task was studied.

**Objective**

The aim of this work is to demonstrate rapid reusability of an existing task-based analytic method for the comparison of the EHR-mediated MedRec task between two different EHR systems used at four different regional perioperative settings.

**Methods**

1. **Clinical Settings**

   This study took place before the enterprise wide conversion to the new EHR. This research involved the following Mayo Clinic settings: Arizona (Phoenix, AZ), Florida (Jacksonville, FL), Methodist and Saint Marys (Rochester, MN). Both Arizona and Florida settings used EHR system 1 (SurgiNet). While both settings at Rochester used EHR system 2 (MICS LastWord) for MedRec.

2. **Data Collection**

   Three investigators collected observational data at the settings across the entire surgical setting (PreOp, IntraOp, PostOp). Data collection and curation was based on the clinical workflow capture and analysis methodology adopted as a standard approach in the Applied Clinical Informatics program at Mayo Clinic.(18)

   As a method of data collection, ethnography entails examining the behavior of the participants in a certain specific social situation and also understanding their interpretation of such behavior.(19) Ethnographic techniques used for this research included interviewing, shadowing, video ethnography, artifact collection, and contextual inquiry. Leadership and key provider
activity interviews were recorded and transcribed. Activity interviews were used to characterize “walk throughs” of key workflows. Shadowing was captured using handheld audio recorders and artifacts were captured using a digital camera. Video ethnography involved capturing HIT interactions with a handheld video camera and with the Morae™ software. Via the use of a webcam, Morae™ was also able to video record participants’ hands, desk space and paper-based artifacts that were used during interactions with the EHR. Morae™ was also used to record audio of participants verbalizing their thoughts and conversations in the immediate vicinity.

One computer scientist specialized in workflow analysis and a computer science graduate student collected the computer logs generated by the EHR at the settings during the same days when the ethnographic study took place. A computer log is a file that records events taking place for a clinical activity during execution of a particular process instance. Computer logs include information about what activity is being performed by whom, when and for which patients. Computer logs in a healthcare setting can be used to discover and study clinical workflows.

3. Data Analysis

Ethnographic videos were analyzed to quantify the duration of the MedRec task and to provide insights on the task.

Programs in Java version 8 were written to automatically detect MedRec activities from computer logs generated by the EHR systems. The discovered computer logs corresponding to MedRec tasks were then imported to Disco™, a licensed tool, for automatic time metric analysis. Descriptive statistics (means and SD) and a time-based histogram, where the bins represent ranges of time duration, were computed in order to identify and eliminate MedRec outliers. Outcomes were validated through triangulation with ethnographic data. A t-test analysis was performed to assess differences in time metrics between sites, with alpha level of 0.5.

Results

1. Data Collection

For all the perioperative sites, a total of 4,400 minutes of video recordings were captured for 99 providers during 16 days. For EHR system 1: at Arizona, 76,000 event logs for 15 providers for a 5–day period were collected; at Florida, 116,706 event logs for 31 providers for a 4–day period were collected. For EHR system 2: across the two Rochester sites, 81,886 event logs for 53 providers for a 7-day period were collected.

2. Data Analysis

2.a. Analysis of video recordings of EHR interactions

Figure 1 presents a schematic representation of the interface used for MedRec at Arizona and Florida. First, the provider navigates to the “Order-Charges” section located on the left side of the interface and clicks on a button labeled as “Document Medication by Hx”. The next screen displays the current medication list and at the bottom a compliance section serves to add the last doses taken. There is no saving option to store changes made.

Figure 2 presents a schematic representation of the interface used for MedRec at Rochester. The process for MedRec involves three steps. First, in the “Outpatient Order-Medication List” page the provider opens a command window and enters a shorthand code “medlist” which directs her to the medications home page. Then the provider selects the “Message” button from the medications home page which takes her to the “Order view” page. There the
provider can leave a message about a medication, reconcile the last dose taken and remove changes to doses. Performing the task of entering medication compliance involves the provider accessing the “Order view” page. The “Add Dose Last Taken” button is clicked which opens the “Dose Last Taken” window for adding the date and time. The “Continue” button is clicked to populate the compliance section with the new date and time. This works as a saving function.

Figure 1: Schematic representations of the MedRec interfaces from EHR system 1 used at Florida and Arizona: A) “Document Medication by Hx” button can be clicked in the “Order-Charges” section, B) the last dose taken of a medication can be added through the “Compliance” section.

Figure 2: Schematic representations of the MedRec interfaces from EHR system 2 used at Rochester: A) the “Message” button can be clicked in the “Outpatient Order - Medication List” page, B) the buttons “Adm/Outpt Verify Message”, “Add Dose Last Taken” and “Clear Dose Last Taken” can be clicked to leave a message about a medication, reconcile the last dose taken or remove changes to doses, C) when the “Add Dose Last Taken” button is clicked it opens the “Dose Last Taken” window for adding the date and time; changes are saved by clicking “Continue”.

2.b. Analysis of computer logs that capture EHR interactions
We manually compared seven video recordings of MedRec tasks performed using the EHR with the corresponding computer logs. This process was used to facilitate the interpretation of the computer logs.

We learned that the computer logs did not have the granularity needed to automatically differentiate between MedRec and other medication-related tasks. We previously discussed how MedRec charting was performed in Arizona and Florida by the providers by accessing the “Order” section. As the Morae™ videos revealed, MedRec was not the only task that could be performed in the “Order” section. Other tasks, like deactivate orders, create a medication list and document medications could be also activated following a similar procedure. Unfortunately, the computer logs did not provide detailed information to differentiate between all those possible tasks and automatically detect time metrics related to the MedRec task.

Table 1 is an example of the sequences of computer logs that represent tasks completed by a provider, Provider1, for a patient, Patient1, at Arizona. The first row shows that Provider1 accessed the “Order” section at 05/04/2016 12:11:23 PM to perform MedRec charting for Patient1, as denoted by Subtimer= ORDERS/PVORDER. At 05/04/2016 12:11:25, Provider1 accessed the “Order” section again, as denoted by Subtimer= ORDERS/PVORDER. While the event logs do not provide information on the specific tasks that were performed, the videos revealed that in the first case Provider1 was performing a MedRec task, while in the second case Provider1 was deactivating orders.

While we provide here one example, we analyzed with the help of video recordings 11 (13.1%) out of the 84 computer log cases available. It was found that it was not possible to automatically differentiate with event logs alone when providers were performing MedRec tasks.

With the help of five video recordings we could determine that computer logs from Rochester contained information to automatically detect MedRec tasks. Table 2 is an example of the sequence of computer logs recorded when Provider2 completed MedRec tasks for Patient2. The first row indicates that Provider2 accessed the screen “RXLIST” at 12/05/2016 02:51:23 PM to perform MedRec for Patient2. The task was completed at 12/05/2016 02:56:45 PM when value of the module column changed from “OCX” to “FLW”. While we can quantify how much time the provider spent during the MedRec task, event logs are not detailed enough to quantify how many medications or to describe the type of medications that were reconciled.

For computer log analysis, 53 perioperative providers at the two Rochester were analyzed. The providers included registered nurses (RN), health unit coordinators (HUC) and certified registered nurse anesthetists (CRNA). A written computer program automatically discovered 383 instances of the MedRec tasks from the collected computer logs. Table 3 summarizes the time metrics calculated by Disco™ from the discovered MedRec instances.

The discovery of MedRec tasks for both Rochester sites through process mining revealed the presence of MedRec instances significantly longer than the mean duration. To automatically remove outliers, we considered all the 383 instances of MedRec tasks and found the Median + 2 standard deviations = 89.25 + 206.2 was approximately equal to 295 seconds.

We use the performance filter of Disco to automatically filter out MedRec tasks that took more than 295 seconds. To further validate the method used to detect outliers, a time-based histogram was created using all the MedRec instances (Figure 3). It was found that 89% of instances had a duration of 300 seconds or less. Time metrics after removing the outliers are also presented in Table 3. Statistically significant differences (p<.001) were founded in MedRec time metrics between sites.
Table 1: Computer logs from EHR system 1 for the MedRec task at Arizona

<table>
<thead>
<tr>
<th>Start time</th>
<th>Activity</th>
<th>Provider</th>
<th>Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>05/04/2016 12:11:23</td>
<td>USR:PWR-SHOW VIEW</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:11:23</td>
<td>USR:REGIMENS.DISPLAY</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:11:24</td>
<td>USR:ORM.LOADORDERPROFILE</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:11:24</td>
<td>USR:POWERPLANS.LOAD PLAN</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:11:25</td>
<td>USR:PWR-CREATE VIEW</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:11:35</td>
<td>USR:PWR-SWITCHFRAME</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:11:38</td>
<td>USR:POWERPLANS.DISPLAY</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:11:42</td>
<td>USR:PWR-SWITCHFRAME</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:11:42</td>
<td>USR:ORM.BUILDS DL</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
<tr>
<td>05/04/2016 12:21:35</td>
<td>USR:PWR-SHOW VIEW</td>
<td>Provider1</td>
<td>Patient1</td>
</tr>
</tbody>
</table>

Table 2: Computer logs from EHR system 2 for the MedRec task at Rochester

<table>
<thead>
<tr>
<th>Start time</th>
<th>Activity</th>
<th>Provider</th>
<th>Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/05/2016 02:51:23 PM</td>
<td>DOT_LSTS RXLIST</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:55:51 PM</td>
<td>RXSTART OCENT</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:55:59 PM</td>
<td>RXPADA OCSEARCH</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:56:02 PM</td>
<td>OCSEARCH EORXNXX1</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:55:49 PM</td>
<td>DOT_LSTS RXPADA</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:56:55 PM</td>
<td>RXSTART RXPADA</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:56:15 PM</td>
<td>EORXNXX1 RXPADA</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:56:21 PM</td>
<td>OCSEARCH EORXNXX1</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:56:18 PM</td>
<td>RXPADA OCSEARCH</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:56:25 PM</td>
<td>OCSEARCH EORXNXX1</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:56:33 PM</td>
<td>EORXNXX1 RXPADA</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
<tr>
<td>12/05/2016 02:56:45 PM</td>
<td>FLWCHT_D FLWADD</td>
<td>Provider2</td>
<td>Patient2</td>
</tr>
</tbody>
</table>

Table 3: Summary of process mining analytics for the MedRec tasks at Rochester

<table>
<thead>
<tr>
<th>Site</th>
<th>Role</th>
<th>Providers</th>
<th>Instances</th>
<th>Total time in sec</th>
<th>Mean (SD) time in sec</th>
<th>Total time in sec</th>
<th>Mean (SD) time in sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saint Marys</td>
<td>RN</td>
<td>21</td>
<td>309</td>
<td>36,058</td>
<td>110.3 (108.3)</td>
<td>273</td>
<td>19,594</td>
</tr>
<tr>
<td></td>
<td>HUC</td>
<td>3</td>
<td>18</td>
<td>9,470</td>
<td>169.1 (91)</td>
<td>45</td>
<td>4,145</td>
</tr>
<tr>
<td></td>
<td>CRNA</td>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>6</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methodist</td>
<td>RN</td>
<td>13</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HUC</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CRNA</td>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>6</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3: Histogram showing frequency of MedRec tasks at two Rochester sites

3. Data Validation
To validate the findings from the analysis of Rochester computer logs (Table 3), first we considered five MedRec instances for which we had both ethnographic videos and computer logs. When we compared their time metrics, we found full agreement.

Second, we used interview data to identify provider’s involvement in the process of MedRec. In Rochester, the RN and HUC roles preformed MedRec while PCA, CRNA roles were not usually in charge of MedRec. Those findings were consistent with the finding from the application of process mining methods, as depicted in Table 3.

Third, we used ethnographic videos and observation outcomes to understand the reasons for the outliers that were automatically discovered and removed. We found in the video’s explanations for the discovered outliers, including:
   a. Interruptions: the provider interacts with the patient, providers, takes phone calls, etc. during MedRec charting.
   b. Use of paper artifacts: the provider writes down certain information regarding the patient.
   c. Multi-tasking: the provider accesses other applications, like medication administration, vital signs charting, etc.

Discussion
A generalizable and reusable mixed method approach was applied to the study of two different EHR-mediated tasks. The proposed analytic approach supports automatic calculation of time metrics from computer logs and removal of outliers using time-based histograms. The method developed for the analysis of the vital sign task was easily reused for the analysis of the MedRec task. Reusability required changing the program so it would automatically identify the computer log signals specific to the beginning and the end of the MedRec task. The rest of the computer-based analytic methods were fully reusable. This led to a significant reduction in the time spent analyzing data, from approximately two months to two weeks.

The method includes validation of findings using ethnographic data and proposes potential explanations for detected outliers. Through data triangulation, this methodology helps to increase the confidence on results and improve the interpretation of the resulting outcomes.

Important lessons were learned from reusing the same mixed-method approach to the study two different clinical tasks. First, the types of research questions that can be answered with automatic methods depend on the quantity and quality of the computer logs used. As it was shown in the case of the prior EHR system used at Arizona and Florida, the EHR did not encode in the event logs the information needed to automatically quantify the start and end of the MedRec task. In contrast, the EHR used at Rochester recorded event logs that could be used to accurately
determine the duration of the MedRec task. However, the event logs collected at Rochester lacked the level of detail needed to identify number and type of medications reconciled in each MedRec session. These findings could be used to guide future decisions on the level of granularity and the type of information encoded in the event logs generated by the new EHR.

Second, there is no standard definition of what constitutes a MedRec task. In its simplest form, MedRec encompasses the creation of a complete and accurate patient medication list, and the comparison of that gold standard list to other available sources. (4) In contrast, the MedRec task could be considered to include the revision of patient’s demographic information, assessment of patient’s medication knowledge, external research on the medication history (e.g. asking the patient’s caregiver), and recommendations for in-hospital prescriptions.(12) These differences in interpretation of what the MedRec task encompasses make it challenging to compare and contrast time outcomes from this study with previous findings.

Mixed method approaches as the one we used here can be time and resource intensive. It requires access to health care environments to support in-situ collection of ethnographic data and collaboration of an interdisciplinary research team with expertise in both computational and ethnographic methods to analyze, triangulate and validate collected data. On the other hand, mixed methods can help build on the strengths of video ethnographic and temporal data mining techniques by integrating the two to overcome the biases inherent in either type of method alone. In the case of ethnography, how generalizable are the findings drawn from small samples? For temporal data mining analysis, how accurate are the discoveries drawn from automatic analysis of big data, when compared with observable data?

An interesting future direction of research will be incorporating into the analysis findings from an automated review of a patient’s EHRs. Similar to computer logs, EHR patient data is already collected by health care facilities. Compared with ethnographic data, the access to EHR data tends to be less resource intensive and intrusive, and more scalable. By accessing a patient’s EHR some of the limitations of the event log analysis could be overcome. For example, in the case of MedRec, EHR data could be used to identify the number of and type of medications that were reconciled for each patient prior to surgery. For another future direction, there are ongoing plans to reuse the task-based mixed method analytic approach explained here, to compare time metrics for key EHR-mediated task’s between pre and post EHR conversion with the goal of informing clinical optimization initiatives at the Mayo Clinic.

Conclusions
This paper demonstrated the rapid reusability of a mixed method analytic approach that combines the use of computer logs and ethnographic data for the study of EHR-based tasks.
These findings offer promising new directions of research. As future work, we plan to further test the applicability of the proposed method to the comparative study of other clinical tasks before and after Mayo Clinic’s conversion to a new EHR. Outcomes of that analysis will be used to inform optimization efforts.

Acknowledgments
Mayo Clinic (BMS0148), for the financial support provided to Arizona State University and Mayo Clinic authors. We also thank the clinicians for allowing us to observe and record their activity. We would also appreciate the help received from Dr. Matthew Burton, Sarah Hirn, Robert Sunday, Sheryl Castaneda and Joanne Bouchereau with ethnographic data collection.
References


Connected Personas: Translating the Complexity of Older Adult Personal Health Information Management for Designers of Health Information Technologies

Dawn K. Sakaguchi-Tang, MS¹, Anne M. Turner, MD MLIS MPH²,³,⁴, Jean O. Taylor, PhD²,³, Julie A. Kientz, PhD¹

¹Department of Human Centered Design and Engineering, University of Washington; ²Department of Health Services, School of Public Health, University of Washington; ³Biomedical Informatics and Medical Education, University of Washington; ⁴Northwest Center for Public Health Practice, University of Washington, Seattle, WA

Abstract

Human-centered design (HCD) can be used to communicate research study findings to designers of health information technologies (HIT). We used the HCD approach to develop personas, scenarios, and design guidelines for designers with the aim that it would lead to new HIT designs that support the autonomy and health of older adults. The foundation of the personas, scenarios, and design guidelines was a study that focused on understanding how older adults manage their health information and the role that stakeholders play in that process. In this paper, we describe how we carried out a HCD approach and how it led us to expand the persona process to create a network of connected personas. The connected personas allowed us to show the complexities of personal health information management for older adults and emphasize the importance of relationships with family, friends, and providers.

Introduction

Personal health information management (PHIM) is a process that includes creating, seeking, organizing, and sharing health information⁵. PHIM for older adults often also includes activities that are supported by family and friends⁶. Health information technologies (HIT), such as patient portals and fitness trackers, have been designed to help people manage their health and health information. However, use and adoption of health technologies by older adults has been slow in comparison to the use of general technologies like email and the Internet⁷,⁸. Human-centered design (HCD) and its methods have been embraced within the field of health informatics and used to develop technologies for patients⁹,¹⁰,¹¹. However, there has been less work addressing the implementation of this approach, especially in regard to translating research study findings into artifacts for designers to use when designing HIT. In this paper, we will describe how we used an HCD approach to develop personas, scenarios, and design guidelines for designers creating HIT for older adults.

Background

HCD, also referred to as user-centered design, is an iterative approach that places impacted people at the center of the design process to ensure that the design and development of a product that is usable, meets people’s needs and goals, and addresses people’s challenges or pain points⁶. In health informatics, HCD has been advocated as an approach that can manage the complexity of health activities, such as the often invisible work of PHIM performed by older adults and their family members and friends¹⁰,¹¹. HCD methods have been used to develop and evaluate health technologies for managing diabetes¹, tracking health indicators¹, and managing personal health records¹².

A crucial goal of an HCD approach is gaining an understanding of people who are impacted by the designs, including end users and other stakeholders. Although this process is generally nonlinear, learning about the people involved—including their context, needs, and goals—is typically placed at the beginning of the process so that this knowledge can play a role in the design ideas and solutions⁶. There are a variety of artifacts that can result from HCD methods. For our project we chose to create personas, scenarios and design guidelines. Personas represent different people and their needs, and provide a sense of their behaviors, attitudes, needs, and goals in specific contexts¹³,¹⁴,¹⁵. Personas are developed from user research, using methods such as interviews, observations, and surveys¹³.

Personas aim to create a common language for designers, to encourage empathy for stakeholders, and to help designers understand stakeholder goals¹³,¹⁴. Personas include elements such as the name of the individual they are describing, a photo of the individual, demographic information, personal goals, pain points (e.g., challenges), and a description of the individual’s use and comfort with technology. Personas often include a scenario or narrative, that illustrates the context surrounding the individual and features their goals, challenges, and experiences with a technology⁴.
Personas have been used and studied in a variety of contexts. In the field of health informatics, personas have been used to raise awareness of a particular population\(^4\) and to translate study findings to inform the design process of health information systems\(^6\). Studies seeking to understand how designers use personas in practice found that personas were often used in the initial stages of the design process as a communication tool within a team, to introduce new team members to their stakeholder groups and to help the team avoid making assumptions about their stakeholders\(^{17,18}\). Personas have also been used to communicate with broader stakeholders, like others within the organization but outside of the project team, about the people using a system\(^{16,19,20}\).

Although the persona method is popular, it has not been without criticism. Some of the main critiques have been that it is difficult to trace persona details to specific data and validate personas using scientific methods\(^{16,21,22}\). In addition, some have argued that the persona method is limited in addressing the complexity of managing health and health information\(^23\). Our study addresses these limitations with a modification of the persona method to accommodate the complexity of PHIM in the context of older adults. As described in this paper, we used findings from an extended study of older adults and their supportive networks to develop an expanded and connected set of personas, demonstrating the web of supportive relationships surrounding a central person’s PHIM processes. This set of connected personas provides a fuller picture of the older adult’s needs and experiences, providing information about the role that family, friends, and providers play in that process. Its overall aim is to improve HIT for older adults to support their health and autonomy. To address these goals, the SOARING study design was based on the Balance Model, an ecological model which takes into consideration the role of social, organizational, technological influences in carrying out specific activities, in this case older adult PHIM\(^24\). As a part of SOARING, we conducted a qualitative study to investigate the health information management practices and needs of adults 60 years and older who lived in a variety of living situations (i.e., independent living, retirement communities, and assisted living). We conducted in-depth interviews with 88 older adults, 52 of their family and friends, and 27 health care providers\(^{25,26}\). Results from this research served as the foundation for the development of the personas, scenarios, and design guidelines.

Methods

This work is a part of a larger project called SOARING (Studying Older Adults & Researching Information Needs and Goals)\(^27\). It is a 5-year project at the University of Washington, funded by the Agency for Healthcare Research and Quality (AHRQ). SOARING seeks to deepen understanding of how older adults manage their health information and the role that family, friends, and providers play in that process. Its overall aim is to improve HIT for older adults to support their health and autonomy. To address these goals, the SOARING study design was based on the Balance Model, an ecological model which takes into consideration the role of social, organizational, technological influences in carrying out specific activities, in this case older adult PHIM\(^24\). As a part of SOARING, we conducted a qualitative study to investigate the health information management practices and needs of adults 60 years and older who lived in a variety of living situations (i.e., independent living, retirement communities, and assisted living). We conducted in-depth interviews with 88 older adults, 52 of their family and friends, and 27 health care providers\(^{25,26}\). Results from this research served as the foundation for the development of the personas, scenarios, and design guidelines.

The persona, scenario, and design guideline development effort was led by a small group of researchers from the SOARING team. We implemented a HCD approach to create these design resources. We refined the personas, scenarios and design guidelines through an iterative process with older adults and designers, as well as through feedback from subject matter experts. Subject matter experts included members of our SOARING team who have expertise in geriatrics, health informatics, nursing, medicine, biostatistics, and human-computer interaction. Figure 1 provides an overview of our HCD approach.

HCD 1.1 Persona Process

We used SOARING study data from the interviews conducted with older adults, family and friends, and providers to draft each set (i.e., older adults, family and friends, providers) of personas. Once we drafted each set of personas (i.e., older adults, family and friends, providers) we sought feedback from our subject matter experts on the accuracy and completeness of the content and the design of the persona. After each feedback session, we incorporated the feedback in an iterative fashion into modifications of the personas. We held at least five feedback cycles for each set of personas - this included in-person meetings with the individual team members and a group of subject matter experts, as well as individual reviews where the personas were sent to the subject matter experts who provided in-document comments. As a result, we created six sets of connected personas.
**Figure 1.** Overview of our human centered design approach

**HCD 1.2 Design Guideline Process**

The aim of the design guidelines was to provide a list of design considerations that would apply specifically to health information technology for older adults and complement the connected personas. To develop the guidelines we drew from the diverse perspectives of HCD designers, older adults, and subject matter experts. We first invited undergraduate and graduate students from the Human Centered Design and Engineering (HCDE) department at the University of Washington to participate in an applied class of developing the design guidelines. We introduced the students to the SOARING project and its findings, as well as to the connected personas. Throughout the quarter, we asked students to brainstorm a list of preliminary design requirements using the personas and through sketching design ideas of health information technologies. Based on this work, students helped to prepare video scenarios of the design requirements to show to older adults in focus groups (Figure 2).

**Figure 2.** Video Scenario. This set of slides illustrates the video scenario of two design requirements that allows older adults with the opportunity to choose who they share their health information with.

We next conducted three focus groups with older adults living in senior living communities and facilities to gain their feedback on the preliminary design requirements. Prior to the start of the session, participants completed a demographic questionnaire. Each session was facilitated by a SOARING researcher and observed by a student or a member of the SOARING team. Three students observed the first focus group, a SOARING researcher observed the second focus group, and one student observed the third focus group. All focus group sessions were recorded. Participants received a handout for each video scenario so they could follow along and take notes. Scenarios were used as discussion prompts to discuss and obtain feedback on preliminary design requirements. After each video,
participants were asked whether the scenario resonated with their experience and their perspectives on concepts such as sharing and privacy.

At study completion, the field notes were coded by a two researchers (DST, MX) using thematic analysis. We used the videos to support our analysis and to verify the themes. We shared the design guidelines draft with the subject matter experts, and modified them based on their feedback. We had at least four feedback cycles.

**HCD 1.3 Study with designers**

We conducted a two-part study with designers to understand their perceptions of the connected personas in comparison to individual personas. We also wanted to understand how designers would use the connected personas in their design process. We conducted two sessions with designers. In the first session, designers worked in groups of 3-5 people and were asked to brainstorm ideas for a design challenge. The design challenge asked designers to design for this question: “How might we help older adults balance receiving support in managing their information related to their health and maintaining independence?”

To compare the designers’ experience between an individual persona and connected personas we used a three phased approach where the groups worked first with just the older adult persona and then were given the connected personas. The designers were then given all six sets of personas consisting of the primary older adult personas and their connected personas. The designers from the first session were invited to the second session. In the second session, designers worked in groups to prototype ideas for the same design challenge from the first session. In addition to the family of connected personas, the designers were provided with the design guidelines to inform their ideas.

All of the sessions were video- and audio-recorded. We also observed the groups and took notes. The groups reported their experiences throughout the design session through feedback forms with open-ended questions. The designers also produced sketches of their ideas and prototypes. Each designer completed an exit survey at the end of every session.

After the sessions, four researchers (DST, KK, AT, YW) conducted a thematic analysis by reviewing the first session video recording. As a group, we reviewed the emerging themes and identified themes code in the rest of the video data. Each researcher reviewed a set of video recordings coded for the themes and transcribed those parts of the video. Then, three researchers used affinity diagramming to organize the themes. One researcher (DST) triangulated the themes with the data from the group discussion feedback forms and individual exit surveys, while being open to additional themes that emerged.

**Results**

The HCD methods we used guided us through an iterative process that was based on study data, informed by feedback from subject matter experts, evaluated by older adults and designers, and supported by academic literature and published information from organizations. This process provided the space for discussions about the complexity of older adults’ PHIM and led to influential decisions about the personas and design guidelines. We discuss below some of the decisions and evaluations that occurred through using a HCD approach, including the emergence of connected personas, outcomes from focus groups that evaluated the design guidelines, and designers’ experiences with both the personas and design guidelines.

**HCD 1.1 Drafting personas and emergence of connected personas**

We started by following the typical persona development process. Our goal was to develop a set of “stand-alone” personas that represented the different experiences of older adults and their family members, friends, and providers. We segmented the older adult personas based upon their living situation because we learned that it had a significant impact on the ways that older adults practiced PHIM. For example, older adults who live in an assisted living facility manage their personal health information differently than older adults who live in a private residence. Older adults who live in an assisted living facility may rely on a variety of staff to gather and store their health information documents (e.g., after-visit summaries), while older adults who live in private residences may keep and file their documents on their own. Other parameters that distinguish the older adult personas from each other included health conditions, their organizational style, and the people in their lives that support their PHIM.

For our first feedback session of the family and friends personas with subject matter experts, the family and friends personas were each drafted as individual personas. Since they were stand-alone personas, the format was similar to the older adult personas. However, as the subject matter experts reviewed these personas and reflected on their
knowledge of older adults and related PHIM research, conversations emerged acknowledging that although the personas were representative of individuals, they did not illustrate the connective nature of the network of people who support the older adults in PHIM. We had gained a deep appreciation of this important aspect of older adult PHIM through embedding the SOARING interviews with family and friends into the ecological framework of The Balance Model\(^4\). This model led us to take an integrated approach to consider older adults’ tasks, tools/technologies, social/organization systems, and the physical environment. Family, friends, and providers played a significant role in the ways that older adults organized, managed and shared their personal health information. Representing the older adults and friends and family in a traditional stand-alone fashion could not adequately capture the rich interconnected nature of these relationships, and the design needs that arose from these connections. As a result, we decided to break from the typical persona method of individual persona groups, and to instead create connected personas.

In order to effectively describe the supportive role of connected personas, we changed the persona layout. The connected persona became one page, with the content focusing on the ways in which each stakeholder supported the PHIM of the older adult they were connected to. Each set of personas features an older adult in a different living situation (i.e., independent residence, retirement community, assisted living, etc.) and connected personas of family and friends and/or providers. Each persona within a set contains a scenario. The older adult scenario describes their current situation in terms of their health, their relationships with the people they are connected to, and with health information technologies, specifically patient portals. Similarly, the scenarios for connected personas (family and friends, and providers) illustrate a viewpoint on their relationship with the older adult and the ways they want to, or do, support the older adult. It also illustrates the goals and challenges they face related to their role in the older adult’s PHIM.

The final result of HCD 1.1 was six sets of personas, with each set featuring an older adult persona as the primary persona and connected personas of family and friends as well as providers who play a role in the older adult’s PHIM (See Figure 3).
The design guidelines were informed by the SOARING interviews, by the personas, and by feedback from older adults and subject matter experts. Below we present themes from the focus groups and feedback from subject matter experts that influenced and shaped the design guidelines. The guidelines encourage designers to consider PHIM-related needs (i.e. issues that might arise as an older adult is managing health information) and user experience needs (i.e. considerations for the use of the health information technology itself). For example, a PHIM-related guideline for family members of an older adult is, Escalate support: Specifically consider tools that will allow older adults to plan for situations in which they may need extra help or full support from family and friends to manage their health and health information. An example of a user experience need is, Design for diversity: Consider approaches like inclusive design to reach people across a wide range situations and with differing abilities. All of the design guidelines were based upon the SOARING findings, academic literature and national organization websites.

To investigate whether our design guidelines resonated with older adults and explore additional design considerations, we conducted three focus groups with older adult participants. In total, 21 older adults participated in the focus groups. The average age among participants was 79 years old. Most participants were female (76%) and a majority identified as white (95%), with a college degree or higher (71%). Most older adult participants reported using a computer on a daily basis (71%) and about half (52%) described their computer experience as being at an intermediate level, meaning between some experience and very experienced. In addition, about half (52%) of participants said that they were using patient portals.

Several themes emerged from these focus groups: keep it simple, support autonomy, balance between maintaining privacy and ease of use, and recognize the diversity among older adults in their use of, and attitudes toward, information technologies. These are described in more depth in the following paragraphs.

Some participants said that they expect that it should be easy to find information and complete tasks using health information technologies. One participant plainly stated, “keep it simple”. This theme also included the ability to understand information, and the suggestion that technologies should avoid using jargon and unfamiliar medical terms. During our sessions, we showed a video that demonstrated features like adjusting size of the text by using icons and buttons. Some participants noted that they were unfamiliar with some of the icons and with some of the terms.

“When I go to my patient portal, it is very simple. That’s what I want. I don’t have to hit all these buttons.” (Participant A, Focus group 2)

“The buttons should be labeled.” (Participant B, Focus group 2)

Another theme expressed was that technology should support autonomy. Several participants resonated with the scenario which described family members requesting access to the older adult’s health information. However, they also expressed that older adults should be able to choose who can have access to their health information. One participant acknowledged the challenge between maintaining autonomy and being safe.
“It’s difficult to maintain autonomy and control privacy and get the help you need. Some people are more willing to give that up than others.” (Participant, Focus group 1)

A third theme centered on the balance between maintaining privacy and ease of use. Ease of use arose as an issue in discussions about logging into a system. Participants expressed their frustrations with remembering passwords. Some participants offered ideas for password solutions, such as using a fingerprint, facial recognition, and voice assistant. However, the importance of privacy also came up in these discussions.

“I object to the idea of using the SSN (social security number), I’m not convinced that it’s safe.” (Participant, Focus group 3)

Although our primary discussions in the focus groups were about the guidelines for HIT, some participants voiced the need to recognize the diversity among older adults in their use of, and attitudes toward, information technologies. Four participants from two focus groups expressed a concern about technology taking away existing services or making those services difficult to use.

“My big concern is that I know a lot of people are like me. They don’t want to use a computer. They don’t want to use a smartphone. They just want to talk to someone. What do we do with those people who cannot get a hold of anybody? There has to be a way.” (Participant, Focus group 2)

In addition to guidelines generated from our own studies, where appropriate we incorporated existing guidelines and best practices for designing for this population and supporting the design of health information technologies. Resources used in the guideline development process include information from published reports, and organization websites like the Pew Research Center, Bureau of Labor Statistics Occupational Outlook Handbook, Center for Disease Control and Prevention (CDC), Alzheimer’s Association, Web Accessibility Initiative guidelines and plainlanguage.gov. We were thus able to address the diversity and range of technology experience among older adults and the importance of involving older adults in the design process. We also used academic literature for creation of the development of the design guidelines. The final result of HCD 1.2 was a set of 37 design guidelines.

HCD 1.3 Assessment: Designers’ experiences with the personas and design guidelines

Our final stage was to provide both the connected personas and design guidelines to a group of designers. There were several themes that came from our two-session study with designers about their experiences with the connected personas and design guidelines. One of the themes was that the connected personas provided designers with a holistic perspective of the older adult and of the problem space. Another was that designers refined their designs to include family, friends, and providers. Although the connected personas were positively received, there were also room for improvement. For example, designers were overwhelmed by the amount of information they had to digest and organize across the six sets of connected personas.

A total of 16 designers participated in the study. In general, the connected personas were well received by designers. A primary benefit of the connected personas expressed by designers was that it provided them with a broader perspective on the older adults’ situation and problem space. It also cultivated empathy for the older adult. They suggested that providing an overview summary of each set of personas with key information, or a video summary about each persona set, would be helpful. They also expressed that the design guidelines were useful and helped them to prioritize their design ideas.

Discussion

This paper outlines our use of a HCD approach to translate and synthesize study findings into design materials in the context of designing HIT for older adult PHIM. We described our development and evaluation of connected personas and design guidelines for designers. Our approach involved close collaboration with SOARING researchers, feedback from subject matter experts, and insights from both older adults and designers. This work supports prior research in health informatics that has translated study data into personas. It also acknowledges the limitations of traditional personas to convey the complexity of health information by challenging the persona method to accommodate the complexity of PHIM for older adults. This work builds on prior work the expands the persona method to accurately describe user groups, their needs and behaviors. Connected personas incorporate the insights we have gained from
taking a holistic approach to investigating older adult PHIM through application of the Balance Model, which considers the individual in the context of their tasks, tools and technologies, environment and organizational structures. Through utilizing this holistic framework, we observed crucial PHIM needs within the connections between older adults and their family, friends, and providers that would not easily be communicated with stand-alone personas. Our personas needed to reflect this complexity because it was essential to the practice of PHIM for older adults. If we had developed traditional personas of individual older adults, we would lose valuable information that only rises out of the relationships between older adults and their social networks.

In using a HCD approach, we found it to be flexible and robust, allowing us to help designers navigate and appreciate the complexities of older adult PHIM. HCD methods have been advocated in health informatics as a way to design and evaluate health information systems that are informed by users, as well as their needs and behaviors. Using a HCD approach provides researchers with a method to translate and communicate complex processes that are critical to the work people do and are necessary for HIT design.

We found that designers appreciated the connected personas and found them to be useful to their design process. We also learned that the connected personas led designers to broaden their perspectives on the users of HIT beyond older adults, to include considerations for the older adults’ family, friends and providers. This finding indicates that the connected personas were successful in communicating the complexity of older adults’ PHIM. Overall, this study demonstrated that using connected personas has the potential to stretch the boundaries of the typical persona method to more accurately represent users and their context. This method may be useful on other complex areas of healthcare in which a variety of roles influence activities, such as care management for children with disabilities and patients undergoing cancer treatment.

Limitations
Our personas and design guidelines are meant to provide information about older adult PHIM that should be considered in the design process of HIT. We did not focus on a particular HIT or explore other activities beyond PHIM. Our results may not be generalizable to other technologies, activities or contexts. We created our connected personas and design guidelines independently from a team of designers. Prior studies have found that designers who are not involved in the development of personas may not use them to make design decisions. Research into the use of these personas in actual practice is needed.

Conclusion
In this paper, we described how using a HCD approach lead to the development of connected personas and design guidelines for designers of HIT to better meet the needs of older adults. This approach allowed us to synthesize study findings about the complexity involved in the ways that older adults manage their health information. In particular, we were able to illustrate the importance of relationships to supporting older adults with their PHIM. Through feedback from designers, we learned that the connected personas were useful in communicating the complexity of PHIM for older adults, and led designers to consider the connections between older adults, their family, friends and providers related to PHIM in their design ideas. Overall, the connected personas and design guidelines helped designers to dive into a greater understanding about older adults and their PHIM. Having this knowledge will lead to HIT that meets the needs of older adults and as a result is used and adopted by older adults.

Acknowledgments: This work was supported by funding from AHRQ #R01HS022106. This material is also based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. (DGE 1256082). The findings and conclusions expressed here are those of the authors and do not necessarily represent the views of AHRQ and NSF. We want to thank Selena Xu and Maria Medina for their hard work in creating the personas. We also want to thank the students who participated in developing the design guidelines as well as Alyssa Bosold for editing this manuscript.

References


An Interprofessional Approach to Clinical Workflow Evaluation Focused on the Electronic Health Record Using Time Motion Study Methods

Jessica Schwartz, RN, BSN1, Jonathan Elias, MD2,3, Cody Slater, BS4, Kenrick Cato, RN, PhD1,3, Sarah Collins Rossetti, RN, PhD1,2

1Columbia University School of Nursing, New York, NY; 2Columbia University Department of Biomedical Informatics, New York, NY; 3NewYork-Presbyterian Hospital, New York, NY; 4Columbia University Vagelos College of Physicians and Surgeons, New York, NY

Abstract

Documentation burden has become an increasing concern as the prevalence of electronic health records (EHRs) has grown. The implementation of a new EHR is an opportunity to measure and improve documentation burden, as well as assess the role of the EHR in clinician workflow. Time-motion observation is the preferred method for evaluating workflow. In this study, we developed and tested the reliability of an interprofessional taxonomy for use in time-motion observation of nursing and physician workflow before and after a new EHR is implemented at a large academic medical center. Inter-observer reliability assessment sessions were conducted while observing both nurses and physicians. Four out of five observers achieved reliability in an average of 5.75 sessions. Our developed taxonomy demonstrated to be reliable for conducting workflow evaluation of both nurses and physicians, with a focus on time and tasks in the EHR.

Introduction

With the integral role that electronic health records (EHRs) play in healthcare today, the implementation of a new system promises to bring numerous changes to both the institution and the workflow of its clinical providers. Quality and clinical outcomes such as preventive screening rates and length of stay are commonly measured before and after an implementation of a new EHR. In addition to potential fluctuations in quality and clinical outcomes, changes in clinician workflow are likely to occur and should be analyzed to address issues and realize potential gains in productivity and/or efficiency. As part of a larger study evaluating the implementation of a new commercial EHR at a large northeastern medical center, we are collecting time-motion data of both nurses and physicians before and after implementation to measure the impact on clinician workflow and documentation burden. Documentation burden can be understood as a combination of many factors, including time, low usability, low satisfaction, and high cognitive spending. Time-motion observations can be used to understand the time component of documentation burden. Collecting time-motion data involves observing a person as they conduct tasks to measure how much time is spent on each task and the sequence of task performance that makes up a workflow. Currently, most time-motion studies that have evaluated clinician workflow in the era of the EHR observed either physician or nursing workflow, but not both. Because one of the overall aims of implementing a new EHR is to reduce documentation burden and increase efficiency for all clinician types, we developed an interprofessional task taxonomy to capture nursing and physician workflow. This taxonomy allows for accurate description of different workflows while obtaining a picture of pooled documentation burden among the patient care team.

While the prevalence of EHRs has grown, documentation burden among clinicians has become a challenge, with time spent on data entry of particular concern. The Office of the National Coordinator (ONC) on Health Information Technology recently released a draft strategy for reducing clinician documentation burden and the first overarching goal is to “reduce the effort and time required to record health information in EHRs for clinicians.” EHR data entry requires that the clinician record clinical findings in the patient’s electronic record, while data viewing is the consumption of entered data. Evidence suggests that both nurses and physicians spend more time documenting after the implementation of an EHR. Increased documentation times are associated with clinician burnout and low satisfaction. In addition to time spent documenting, poor alignment of the EHR with optimal care delivery workflows is a known challenge. The ONC names “better alignment of EHRs with clinical workflow” as the first strategy to improve EHR usability. As we work to support healthcare’s Quadruple Aim – improving the care
experience, improving population health, reducing cost, and improving the well-being of providers – it is important to understand the role that EHRs play in either hindering or aiding in providers’ care delivery\textsuperscript{19,20}.

Much research has been conducted to measure documentation burden and the role of EHRs in clinician workflow\textsuperscript{4,6–14}. While many investigators are utilizing EHR log file data to measure documentation times, a limitation to this methodology is lack of validation by direct observation\textsuperscript{14}. For example, while the log files may indicate that a user spent thirty minutes of uninterrupted time viewing a note, it is possible that the user walked away while remaining logged in and delivered direct patient care or was interviewing the patient while reading the note. This patient-care time would be inaccurately reported as note-viewing time only, without the utilization of direct observation. Time-motion observation serves as a method for corroborating findings from EHR log data and for incorporating tasks that occur outside of the EHR.

Much of the current literature surrounding EHR documentation burden is centered on physician documentation burden and has offered strategies involving redistribution of workload from physicians\textsuperscript{2,10,21,22}, including adding order entry to the responsibility of “clinical staff”\textsuperscript{10}. Such efforts should include consideration of existing documentation burden among other clinical roles, such as nursing. Studies have shown that nurses are also dissatisfied with the current state of EHRs and cite EHR inefficiencies as barriers to delivering best-practice patient care\textsuperscript{3,23}. A potential unintended consequence of restricting EHR burden analysis to one clinical role is that low-satisfaction tasks like data entry will be passed from one role to another and improvement will not occur across the care team. By utilizing the same tasks while observing both nurses and physicians, we seek to understand the documentation burden of both clinical roles within the care team, as well as how they are both impacted after a new EHR is implemented.

**Methods**

In July and August of 2018, the primary author conducted a review of the literature on time-motion evaluations of clinician workflow, specifically surrounding the use of electronic health records (EHRs). We did find that Ballerman and colleagues utilized a shared taxonomy to conduct time-motion observation of both nurses, physicians, and respiratory therapists, though their study was conducted before an EHR was implemented, thus did not have EHR-specific task names\textsuperscript{24}. Therefore, we could not utilize their taxonomy for our purposes. Task names were extracted from relevant articles and brought to an interprofessional team of clinical researchers (RNs and MDs)\textsuperscript{4,6–9,25,26}. Each task was evaluated for its relevance to nursing and physician workflow and then re-worded and re-defined in order to accurately describe the work executed by each profession. For example, in a previous study conducted by Yen and colleagues, evaluating nursing multi-tasking on a medical surgical unit, a task named “Direct – Procedure” was defined as “RN performs treatment or procedure that cannot be delegated (top of license task)”\textsuperscript{24}. We redefined this task so that procedures performed by both nurses and physicians could be captured under “Direct Care – Procedure”. Our task definition is, “Performs direct procedure (e.g. wound care, blood draw, CVC placement)”. Therefore, when a nurse performs wound care or phlebotomy, or a physician places a central venous catheter, our observers capture this action as a “Direct Care – Procedure” task. (Note: we are not observing clinicians in the operating room, therefore operating procedures are out of scope). With this approach, we can capture a pooled measurement of time spent performing procedures by both nurses and physicians before and after implementation, with respect to the differing types of procedures they are each licensed to carry out. EHR tasks such as data entry also often differ between nurses and physicians. Physicians typically enter their clinical findings into a note, while nurses may spend more time entering findings into flowsheets or care plans and synthesize in shorter “end of shift” notes. Instead of employing multiple tasks, we capture each activity in our definition of “Data – Entering.” This allows for measurement of data entry across the two roles, as well as increases the likelihood that observers will be able to reliably capture each of these tasks because they have fewer tasks to choose from.

The task list evolved over a period of four months through team discussion, lab training, and clinical-setting training. After task alterations were made to accommodate nursing and physician workflow, any further alterations were made to reach a level of granularity that allowed us to reliably capture tasks in real time. The taxonomy was pared down from 43 to 38 final tasks, including “tasks” that capture the clinician's location and communication activities (Table 1). Efforts were made to prioritize capture of data entry tasks, data consumption tasks, and patient care tasks, because much of the current sentiment around EHRs reflects frustration at their functional role as time-consuming data repositories, detracting from patient care, rather than as sources of clinical information delivery\textsuperscript{3,10,15,16}. The taxonomy is organized into the three overarching categories – location (where the clinician is, physically), task (what the clinician is doing), and communication (conversations the clinician is having or listening to). These categories were developed and validated for time-motion evaluations by investigators at The Ohio State University Department of Biomedical Informatics in their creation of a web-based time-motion observation platform, TimeCat\textsuperscript{27}. TimeCat allows the
observer to document up to one active task in each category at any time during the observation so that the clinician’s location, what they are doing, and any conversations they may be having, simultaneously, are recorded. TimeCat version 3.9 was used in this study.

**Table 1.** Final taxonomy

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>Data – Entering</td>
<td>Writing any clinical note in EHR (e.g. progress note, procedure note, admission note, discharge note, result note in a lab test). Writing note or letter in EHR messaging system. Documenting in flowsheets, forms, admission/discharge navigators, patient education, plan of care, allergies, problem list, visit diagnoses, or other data capture functionality.</td>
</tr>
<tr>
<td>Task</td>
<td>Data – Viewing</td>
<td>Viewing information within notes, handoff/sign-out tab, flowsheets, problem lists, demographics, past medical/family history, results section, admission/discharge navigators, patient education, plan of care, reports, or other type of data display or visualization.</td>
</tr>
<tr>
<td>Task</td>
<td>Data – Viewing Patient Data Archive</td>
<td>Viewing information in patient data archive application window.</td>
</tr>
<tr>
<td>Task</td>
<td>Direct Care – Physical Assessment/Exam</td>
<td>Performs direct physical assessment/exam.</td>
</tr>
<tr>
<td>Task</td>
<td>Direct Care – Procedure</td>
<td>Performs direct procedure (e.g. wound care, blood draw, CVC placement).</td>
</tr>
<tr>
<td>Task</td>
<td>Email (Desktop)</td>
<td>Sends email outside of EHR at desktop or laptop computer (not on a mobile phone).</td>
</tr>
<tr>
<td>Task</td>
<td>Handoff/Sign-out – Documenting</td>
<td>Adding notes and tasks to list of treatment/care goals for which the oncoming clinician will be responsible. Using EHR handoff or sign-out activity.</td>
</tr>
<tr>
<td>Task</td>
<td>Indirect Care</td>
<td>Activities associated with equipment search, arranging the unit, changing the patient to another bed within the same unit, “waiting time” to perform other tasks, or searching for patient, clinician, or other resource.</td>
</tr>
<tr>
<td>Task</td>
<td>Log into EHR</td>
<td>Logs in/signs in to the EHR system.</td>
</tr>
<tr>
<td>Task</td>
<td>Log out of EHR</td>
<td>Logs out/signs out of the EHR system.</td>
</tr>
<tr>
<td>Task</td>
<td>Med Administration</td>
<td>Barcode or non-barcode medication administration, medication infusion titration, scanning patient and medication/vaccine and completing documentation in EHR medication administration record (mar) or immunization activity.</td>
</tr>
<tr>
<td>Task</td>
<td>Med Preparation</td>
<td>Obtaining medications from medication dispensing system, opening packages, priming IV tubing, flushing line.</td>
</tr>
<tr>
<td>Task</td>
<td>Med Reconciliation</td>
<td>Discontinuing and adding new medications in EHR medication reconciliation activity.</td>
</tr>
<tr>
<td>Task</td>
<td>Smartphone Clinical Messaging App</td>
<td>Non-phone-call use of designated clinical smartphone phone, (including messaging, reference lookup, or EHR app).</td>
</tr>
<tr>
<td>Task</td>
<td>Mobile Phone</td>
<td>Non-phone-call use (e.g. texting or sending email), excluding designated smartphone clinical messaging app.</td>
</tr>
<tr>
<td>Task</td>
<td>Orders – Entering</td>
<td>Orders entered into EHR order entry activity.</td>
</tr>
<tr>
<td>Task</td>
<td>Pager – Sending/Viewing Msg.</td>
<td>Viewing or sending message including through EHR paging or web-based paging.</td>
</tr>
<tr>
<td>Task</td>
<td>Paper Chart/Notes – Documenting</td>
<td>Writing on any paper components of the patient's medical record (e.g. notes, letters, results, consent) or a printed list/schedule.</td>
</tr>
<tr>
<td>Task</td>
<td>Paper Chart/Notes – Viewing</td>
<td>Viewing any paper components of the patient's medical record (e.g. notes, letters, results, consents) or a printed list/schedule.</td>
</tr>
</tbody>
</table>
Table 1. Final taxonomy continued

<table>
<thead>
<tr>
<th>Task</th>
<th>Patient List/Schedule – Viewing</th>
<th>Viewing EHR clinic schedule or patient lists of unit or service.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>Personal</td>
<td>Eating, personal computer use, using restroom.</td>
</tr>
<tr>
<td>Task</td>
<td>Reference Materials – Viewing</td>
<td>Using medical/clinical reference site (e.g. UpToDate\textsuperscript{28}), including accessing info-button through EHR.</td>
</tr>
<tr>
<td>Task</td>
<td>Transcribing</td>
<td>Taking notes onto paper from EHR.</td>
</tr>
<tr>
<td>Task</td>
<td>Travel</td>
<td>Moving from one area to another.</td>
</tr>
<tr>
<td>Task</td>
<td>Use of Other CIS</td>
<td>Logging into another clinical information system (CIS) besides the EHR or viewing another system that does not require log in (e.g. telemetry).</td>
</tr>
<tr>
<td>Communication</td>
<td>Handoff/Sign-out</td>
<td>Discussing to-do list or plan of care with oncoming clinician, including events in previous shift and outstanding/upcoming goals.</td>
</tr>
<tr>
<td>Communication</td>
<td>Phone Talking</td>
<td>Conversation over phone, either in team area or on mobile device.</td>
</tr>
<tr>
<td>Communication</td>
<td>Rounding and Meetings</td>
<td>Planned uni- or interprofessional discussion of patient treatment and care goals.</td>
</tr>
<tr>
<td>Communication</td>
<td>Code/RRT</td>
<td>A code team / rapid response team (RRT) meeting has been initiated for a patient</td>
</tr>
<tr>
<td>Communication</td>
<td>Verbal w/Patient/Family</td>
<td>Non-rounding discussion with patient/family (e.g. patient education, shared decision making, or live communication via phone interpreter).</td>
</tr>
<tr>
<td>Communication</td>
<td>Verbal w/Staff – Care-Related</td>
<td>Non-rounding discussion, not including family meetings. Includes non-rounding presentations to physicians.</td>
</tr>
<tr>
<td>Communication</td>
<td>Verbal w/Staff – Not Care-Related</td>
<td>Discussion that is not about patient or patient care.</td>
</tr>
<tr>
<td>Location</td>
<td>Hallway</td>
<td>Corridor in clinic or on unit between patient/exam rooms and team areas.</td>
</tr>
<tr>
<td>Location</td>
<td>Inaccessible Patient Room</td>
<td>Isolation room, patient under security, patient asked not to be observed.</td>
</tr>
<tr>
<td>Location</td>
<td>Patient Room</td>
<td>Inpatient room or outpatient exam room.</td>
</tr>
<tr>
<td>Location</td>
<td>Supply Room/Medication Administration Room</td>
<td>Designated room or area containing clinical supplies or where medications are administered (clinic).</td>
</tr>
<tr>
<td>Location</td>
<td>Team Area</td>
<td>Nurses station, designated huddle area, clinician workroom.</td>
</tr>
<tr>
<td>Location</td>
<td>Waiting Room</td>
<td>Open area where patients/family wait to be seen.</td>
</tr>
</tbody>
</table>

*Viewing historical patient data will happen in this application after go-live of new EHR

After obtaining IRB approval, we evaluated the interprofessional taxonomy by conducting inter-observer reliability assessments (reliability sessions) on an acute care/step-down unit. The five observers in our study include two licensed and experienced registered nurses, two licensed and experienced medical doctors, and one medical student. Reliability sessions involved two observers following the same clinician, either a nurse or a physician, simultaneously documenting tasks conducted for one and a half to two hours. All activities, clinical and non-clinical, were observed throughout the session and observers entered patient rooms unless the patient was on isolation precautions or the patient or clinician requested otherwise. Each clinician consented to being observed before the session began. Observers did not converse with each other or the clinician during the observations and the data were compared for similarity afterward.

TimeCat includes a sophisticated module for conducting reliability sessions which displays similarity scores in domains called Proportion, Naming, Duration, and Sequence\textsuperscript{4,29}. It also displays a side-by-side visualization of the tasks captured by each observer throughout the observation (Figure 1). Each of the domains were developed and validated by Lopetegui and colleagues and are described in a publication pending review\textsuperscript{29}. The Proportion domain assesses the proportion of time that the two observers document that a specific task is occurring\textsuperscript{29}. For example, if in 10 seconds, Observer 1 documents that the clinician is entering orders, and during 8 of those seconds Observer 2 documents that the clinician is entering orders, the Proportion agreement would be 0.8. This data is then used to calculate Cohen’s kappa agreement score. The Naming domain evaluates each task and compares its name to the task name with which it has the most time overlap\textsuperscript{29}. By measuring the proportion of tasks that share the most time overlap...
that also share the same name, this score reflects observer agreement on the task that is occurring at any given moment. This domain is also measured with a Kappa agreement score. The Duration domain builds off the Naming domain by calculating a concordance correlation coefficient from tasks with the most time overlap, to reflect the agreement between observers on the length of the task. The Sequence domain reflects agreement between observers on the sequence of tasks, no matter their duration. This is measured with a Needleman-Wunsch algorithm score which was developed using informatics approaches to sequence amino acids. Each domain and its score definition is outlined in Table 2 and an example of the output in TimeCat is shown in Figure 1. Domain scores are calculated for each of the three categories – Task, Location, and Communication.

Table 2. Inter-observer reliability domains, measures, and definitions from Lopetegui et al.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>Kappa</td>
<td>Agreement that a specific task is occurring</td>
</tr>
<tr>
<td>Naming</td>
<td>Kappa</td>
<td>Agreement on the name of tasks that have the most time overlap</td>
</tr>
<tr>
<td>Duration</td>
<td>Concordance Correlation Coefficient</td>
<td>Agreement on the duration of tasks sharing the most time overlap and same name</td>
</tr>
<tr>
<td>Sequence</td>
<td>Needleman-Wunsch</td>
<td>Agreement on sequence of tasks</td>
</tr>
</tbody>
</table>

Our criteria for achieving reliability is that within a minimum of three observations, the observer must obtain at least two with a score of at least 0.8 in the Proportion and Duration domains, a score of at least 0.7 in the Sequence domain, a recent observation scoring at least over 0.65 in each, and with demonstration of improvement over time. This definition is consistent with previous work delineating substantial reliability between observers. Once an observer achieves reliability, they become a “gold standard” for training future observers and can begin conducting observations on their own. Because many time-motion studies measure and report inter-observer reliability inconsistently, Lopetegui and colleagues outlined suggestions for setting inter-observer reliability goals according to the research question being asked, rather than aiming to achieve high scores in each domain. Following these guidelines, we determined that the Proportion, Duration, and Sequence domains are most important for our reliability sessions. Our aims are two-fold; first, to understand how much time nurses and physicians spend entering data, viewing data, and caring for patients, and second, to capture a picture of their workflows. The
Proportion and Duration assessments test our ability to reliably meet the first aim by measuring agreement on task occurrence and duration. In addition to these assessments, the Sequence domain tests our ability to reliably meet the second aim by measuring agreement on the order of tasks. These scores, therefore, take priority over the Naming domain, which is of more importance for studies interested in answering the question, “What do clinicians do at any given time?”29, which is not a target of our study.

Results

Twenty reliability sessions were completed. All sessions were conducted between 7am and 7pm, at varying times during this window. Three sessions were conducted on Monday, five on Tuesday, one on Wednesday, five on Thursday, and six on Friday. Four out of five of our observers achieved reliability in as few as four reliability sessions and as many as eight (average 5.75 sessions/observer). One of our sessions included perfect agreement on Communication Proportion, Communication Duration, and Communication Sequence. Three included perfect agreement on Location Duration and one included perfect agreement on Location Sequence.

Additionally, inter-observer reliability scores trended positively, regardless of whether the observer was following a nurse or a physician. For example, two of our observers achieved one reliable observation while observing a nurse, and then their second reliable observation while observing a physician. It might be hypothesized that the nurse observers would be biased toward a nursing interpretation of each task and vice versa for the physicians/medical student. However, our taxonomy demonstrated to be reliably executable by nurses, physicians, and a medical student while observing both nurses and physicians on an acute care hospital unit. Table 3 displays the results of our reliability sessions.

Table 3. Inter-observer reliability results by observer

<table>
<thead>
<tr>
<th>Observer</th>
<th>Reliability Sessions</th>
<th>Communication Range</th>
<th>Task Range</th>
<th>Location Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PK¹</td>
<td>CCC²</td>
<td>NW³</td>
</tr>
<tr>
<td>1 (RN)</td>
<td>6</td>
<td>0.43-1</td>
<td>0.155-1</td>
<td>0.62-1</td>
</tr>
<tr>
<td>2 (RN)</td>
<td>5</td>
<td>0.4-0.81</td>
<td>0.367-0.988</td>
<td>0.71-0.82</td>
</tr>
<tr>
<td>3 (MD)</td>
<td>8</td>
<td>0.6-1</td>
<td>0.682-1</td>
<td>0.68-1</td>
</tr>
<tr>
<td>4 (Medical Student)</td>
<td>4</td>
<td>0.76-0.97</td>
<td>0.155-1</td>
<td>0.72-0.86</td>
</tr>
<tr>
<td>5 (MD)</td>
<td>Not yet achieved</td>
<td>0.4-0.81</td>
<td>0.367-0.979</td>
<td>0.62-0.8</td>
</tr>
</tbody>
</table>

*Number of inter-observer reliability sessions the observer needed to conduct before meeting reliability criteria.
PK¹=Proportion Kappa, CCC²=Duration Concordance Correlation Coefficient, NW³=Sequence Needleman-Wunsch Score

Figure 2 shows a task time agreement matrix for a reliable session. In the agreement matrix, tasks captured during the session are listed across the top of the grid and along the left of the grid. The number of seconds that both observers agreed on the task occurring is plotted across the grid in green. The numbers in red display the seconds that the observers disagreed on the occurring tasks and they are plotted at the intersection of the tasks documented by each observer.

Figure 3 shows a plot of task duration correlation for a reliable session. In this plot, each task, designated by a colored dot, is plotted with an x-value denoting the number of seconds Observer 1 documented that the task occurred, and a y-value denoting the number of seconds Observer 2 documented that the task occurred. A 45-degree line is plotted as
a reference point of perfect duration agreement. TimeCat generates these plots and labels observers as either “Gold standard” or “Trainee” to reflect the inter-observer reliability process of validating new observers.

**Figure 2.** Example task time agreement matrix (generated by TimeCat)

**Figure 3.** Example task duration correlation (generated by TimeCat)

**Discussion**

The outcomes of our reliability sessions are significant because they demonstrate that with careful consideration of task taxonomy and adequate training, observers from differing backgrounds can reliably capture tasks performed by a nurse using the same taxonomy as tasks performed by a physician, and vice versa. Additionally, our specific taxonomy has demonstrated to reliably encompass workflow among nurses and physicians using an EHR. Also, by conducting our inter-observer reliability sessions over varying times and days of the week, we can be confident that we tested reliability during the work hours in which we will collect real observation data. Future studies looking to evaluate shared documentation load and to capture a picture of clinician workflow may model this approach and utilize this taxonomy.

All care team members should be considered when looking to reduce EHR burden and improve alignment of the EHR with clinicians’ workflow. The interprofessional taxonomy developed here was shown to be reliable in our settings for conducting workflow observations of both nurses and physicians to answer research questions surrounding EHR tasks and their duration time. Further, our taxonomy proved to be reliably executable by observers with both nursing and medical backgrounds. Though one of our observers has not yet achieved reliability, that observer is receiving further training to help meet criteria.

The level of detail in our taxonomy allows us to reliably answer questions surrounding changes to time spent on patient care, data entry, and data viewing, as well as changes to general workflow. Using this data to accompany more granular detail of user behavior from EHR audit log data will help us identify areas where efficiency and/or workflow alignment is gained from the new EHR, as well as areas for improvement, for both nurses and physicians. We will be employing this taxonomy to conduct time-motion observations in both the hospital and ambulatory settings. Future work may expand the use of this taxonomy to study additional clinical roles such as social workers, care coordinators, and therapists. Incorporating even more members of the care team will further strengthen our understanding of shared documentation burden and the role of the EHR in clinicians’ workflow. Additionally, this time-motion and EHR audit log data could be used to quantify the care team documentation time required per patient.
Limitations

Though time-motion study of workflow is an ideal accompaniment to EHR audit log analysis, we acknowledge that it can be prohibitively resource intensive, and thus feasibility in all settings is limited. In consideration of resources and scope, we are only focused on day shift hours Monday through Friday, though we acknowledge that night shift and weekend workflow should be considered in future work. We also acknowledge that our task taxonomy might not be useful for answering questions such as, “How much time do clinicians spend multitasking during note documentation?” As stated previously, we found it prudent to sacrifice capturing this level of detail in our time-motion analysis in favor of reliably capturing the broader workflow. We plan to utilize the audit log data to analyze time and click navigation patterns surrounding EHR tasks such as note documentation. However, more granular time-motion studies may be conducted to focus on specific EHR tasks with taxonomy modification. Additionally, this study is primarily focused on evaluating the workflows of nurses and physicians, though, as mentioned previously, there are certainly other care team members integral to patient care who could benefit from a similar inclusive approach to workflow analysis using and extending this taxonomy, as needed.

Conclusion

In summary, we were able to validate an interprofessional taxonomy that is useful for reliable EHR-focused time-motion observation of nursing and physician workflow. Our work is significant because it demonstrates the potential to utilize one method to quantify burden and analyze workflow across different clinical roles, at a time when reducing EHR burden and improving EHR congruence with clinical workflow is a national priority. If the lens isn’t widened to examine more than one member of the care team when looking to improve the role of the EHR in clinicians’ care, then myopic strategies may be implemented that do not ultimately improve care delivery because some clinicians will continue to be inefficient and overburdened. By capturing the influence of EHRs among multiple care team members’ workflows and burden, we can focus on visionary re-design strategies of EHR tasks to promote care efficiencies that ultimately aid patient-centered care. This taxonomy can be used in future work looking to evaluate the role of the EHR in clinical workflow and documentation burden so that improvements can be made for all members of the care team.

Acknowledgements

Jessica Schwartz is a Pre-Doctoral Fellow funded by the Reducing Health Disparities through Informatics (RHeaDI) training grant funded by the National Institute for Nursing Research (T32NR007969). We would also like to thank the nurses and physicians who participated in this project.

References

7. Mamykina L, Vawdrey DK, Hripcsak G. How do residents spend their shift time? A time and motion study


28. 2019 UpToDate I. Evidence-Based Clinical Decision Support at the Point of Care | UpToDate [Internet]. [cited 2019 Mar 4]. Available from: https://www.uptodate.com/home


Effects of a Virtual Pointer on Trainees’ Cognitive Load and Communication Efficiency in Surgical Training

Azin Semsar, MS¹, Hannah McGowan¹, Yuanyuan Feng, MS¹, Hamid R. Zahiri, DO ², Ivan M. George ², Timothy Turner, PhD ², Adrian Park, MD², Helena M. Mentis, PhD¹, Andrea Kleinsmith, PhD¹

¹University of Maryland, Baltimore County, Baltimore, Maryland, US; ²Anne Arundel Medical Center, Annapolis, Maryland, US

Abstract

We investigated the cognitive load effect of a trainer providing surgical instruction by pointing/drawing over laparoscopic video to a trainee. Results showed that while cognitive load is higher overall with the use of the instructional system, there is a decrease by the second experience of being instructed by the Virtual Pointer. Further analysis showed that trainees were more likely to perform the surgical task and watch/listen to the trainer’s instruction at the same time when the instructional system was used. This is thought to be an indication of more efficient communication when using the instructional system. Thus, although there is a small cognitive overload with the instructional system initially, the more efficient communication allows trainees to better integrate the knowledge and instructions being conveyed into the actions they must perform - indicating a better learning environment.

Introduction

Evidence demonstrates that both cognitive underload and overload leads to decreased performance ¹,² and learning ³,⁴. Greater learning outcomes are predicted when training tasks and technologies that reduce cognitive load are employed. The reasoning is simple – more working memory resources will be available for learning ⁵,⁶. Understanding the effects of cognitive load in medical training is especially important given that many of the necessary activities require the learner to integrate different skills, knowledge and behaviors concurrently ⁷. Studies have shown that trainees have fewer cognitive resources, i.e., less working memory, available for decision making when comprehension of the instruction is more demanding ⁸.

To evaluate the effects of instructional technologies on cognitive load, the multimedia dual-processing theory ⁹ may be a consideration. This theory posits that there are two systems for processing information, one for visual information and another for audio/verbal information ¹⁰,¹¹. Both systems have a finite amount of processing capacity, thus the presentation of two forms of visual information require the visual processing system to be divided in two, leaving the audio processing resources untouched. The claim is that new technologies that can make use of both processing systems allow for more availability of the total cognitive resources ¹²,¹³. Studies investigating this theory have demonstrated greater knowledge acquisition is achieved, and cognitive resources are not overloaded when both processing systems are used, i.e., the audio/verbal channel and the visual channel ¹⁴,¹⁵,¹⁶.

We employed both subjective and objective measures of cognitive load of surgical trainees while performing tasks and being instructed by a trainer using an instructional technology called The Virtual Pointer. The Virtual Pointer enables a trainer to point or draw a free hand sketch over live laparoscopic video; adding the visual instruction channel to the existing traditional verbal instruction. Adding an additional channel of communication in a dyadic interaction can have positive implications for communication efficiency ¹⁷. Cognitive load can be influenced by communication efficiency in a training interaction ¹⁸. The more efficient the communication, the less cognitive processing is required for understanding the communicated content. Therefore, more cognitive resources are available for learning the content. Our hypothesis was that a Virtual Pointer can enhance communication efficiency, and consequently allow for further cognitive processing.

Related Work

Cognitive Load And Information Processing Channels

Cognitive load theory ¹⁹,²⁰,²¹,²² describes working memory as having a limited capacity in processing information. When the number of informational units exceed the limited capacity of working memory, new information cannot be processed and thus learning will be hindered. In a learning environment, cognitive load may arise from the instructional tools and techniques rather than the complexity of the task itself. Learning, as an outcome of a training process, will be hindered if instructional materials and methods overwhelm a trainee’s cognitive resources ²⁰. In
multimedia learning (i.e., learning from words and pictures), cognitive load arising from instructional techniques is a central challenge faced by designers of multimedia instructions. Mayer and Moreno (2003) suggest nine ways to reduce cognitive load in multimedia learning including a cognitive process involved in integrating verbal and visual cues. One other recommendation they make is to off-load one information receiving channel (e.g. verbal and visual) by splitting the information among multiple channels. This solution reduces the processing demand of one channel so the learner has further capacity for processing within other channels. They demonstrated that students understood a multimedia explanation better when the words were narrated rather than shown as an on-screen text.

Brunken et al.’s study confirms the guideline of splitting the information among multiple channels using a dual-task methodology. In their study, participants were asked to perform two tasks at the same time: the primary task involved learning from a multimedia program on how the human cardiovascular system works, and the secondary task involved needing to press a button as soon as a letter color changed. Analyzing the reaction time for the secondary task as a cognitive load measure, they found that reaction times were lower for the audiovisual presentation compared to the visual only presentation. This showed that a combination of audio and visual presentation induced less cognitive load than a single modality presentation, i.e visual-only. The dual-task methodology is based on the assumption that cognitive capacity is limited, but can be flexibly allocated. In the case of processing two tasks at the same time, the cognitive resources have to be split between the two. Therefore, when a person is able to better perform a second task simultaneously with the first task, the less cognitively loaded the person is on either task, thus allowing for better integration of cognitive processing.

To conclude, the way instructional techniques and tools are constructed may affect cognitive load differently. While it is suggested that designing an instructional method in a way that splits the information between auditory and visual channels reduces the cognitive load, a Virtual Pointer has not been studied in terms of how it affects the cognitive load even though it uses the same principle of splitting information between two channels. Because it has been reported that multiple modalities of processing channels allows for more cognitive processing space, we are interested in examining if the Virtual Pointer may affect cognitive load for this reason.

**Cognitive Load and Communication**

Analyzing cognitive load in communication settings is additionally important to allow for effective group collaboration. For example in Computer Support for Collaborative Learning (CSCL), it has been discussed that the interaction among subjects in a learning environment may generate communication activities (e.g. explanation, disagreement, and mutual regulation), which may trigger extra cognitive mechanisms (e.g. knowledge elicitation, and internalisation). The primary reason for the importance of cognitive load in CSCL environments is that CSCL environments force learners to coordinate one or more external, instructional representations. In such environments, the need for learners to integrate the textual, verbal, and visual information sources presented to them creates an additional cognitive load on top of the task itself. The balance between the reduction of individual computing due to division of labor and the increase of individual computing necessary for interaction is the key to avoid detrimental cognitive overload in CSCL environments.

Effective instructional methods can enhance the communication efficiency between the learner and the trainer. Once the communication becomes more efficient, the cognitive resources will be directed toward activities that are relevant to learning. Communication is a joint activity, which is coordinated based on common ground – shared knowledge, beliefs, and suppositions. And in turn, common ground is incrementally built on the previous joint activities within a group. Thus, as common ground accumulates, communication becomes more efficient and efficient communication reduces the costs for the development of common ground.

In conversation, one major coordination task is turn taking, i.e., all people conversing coordinate the time of entry and exit in the conversation. The changes in the coordination of turn taking on one hand indicates the efficiency of communication, and on the other, relates to the grounding costs, such as the costs in language processes, i.e., the construal of meaning, and the costs in signaling and accepting. For example, more turns, fewer words and more synchronicity manifest in teams with an increased amount of shared understanding, and thus more efficient communication. The increased common ground and improved communication efficiency facilitates the construal of what would be presented, manifested by more quick and short turns. For instance, Fussell et al. studied the effects of shared visual context in a collaborative repair task on communication efficiency through turn taking analysis. The greater number of turns in the audio-video instruction compared to audio-only instruction shows that audio-video instruction enhances the communication efficiency between the trainee and trainer in performing a repair task. Because of this, we decided to not only investigate cognitive load, but communication efficiency elicited by the Virtual Pointing system because of its integrated audio-visual instruction basis.
Cognitive Load in Surgical Training

An investigation of the effect of haptic feedback on surgical residents’ cognitive load when performing a laparoscopic task in a virtual reality setting while they were imposed by a secondary mental arithmetic task showed that residents performed 36% faster with haptics compared to without. This suggests that the addition of haptic technologies allowed for a reduction in cognitive load and thus increased performance efficiency. In comparison, Andersen, et al. found that adding technology to a surgical training condition neither increased or decreased cognitive load nor did it affect performance. Here, they compared trainees’ cognitive load with and without a simulator-integrated tutor function in a virtual reality surgical setting by analyzing the reaction time of secondary observational task. They found the integrated tutoring did not influence reaction times and did not have an effect on cognitive load. They did, however, observe that novices gain proficiency in the Virtual Reality surgical simulation after relatively few practice sessions. This effect of experience, i.e., increase in proficiency after additional exposure to using a new technology, has been observed in other studies and could play a critical role when examining a technology’s effect on cognitive load. For example, Theodoraki et al. assessed surgeons’ cognitive load of using an image-guided navigation system - a system that combines the information captured from cameras, ultrasonic, electromagnetic sensors and relay the patient’s body view and the surgeon’s movements in relation to the patient, to the surgeon’s screen. Non-significant difference in heart rate and heart rate variability between the two conditions showed that cognitive load did not differ between use with and without the navigation system. Nevertheless, the heart rate variability was slightly higher in the navigation-supported condition. Moreover, they observed that residents who had more practice in performing the procedures showed a slight decrease in mental workload while using the navigation system. This suggests that with more experience, there may be more cognitive processing space open to allow for the addition of new technology to be beneficial.

Material and Method

System Design and Setup

The Park Trainer (Stryker Corporation, USA) was used for the simulated laparoscopic tasks (Figure 1.c). It consists of a housing unit for physical anatomical models, a flexible shield with openings for the laparoscopic camera and instruments to be inserted, a standard laparoscopic camera with light source using the Stryker computer system, and a standard laparoscopic monitor on an adjustable arm at the top. The Virtual Pointer was designed to facilitate the conveyance of knowledge during surgery by enabling attending surgeons to point or draw on the laparoscopic video for a surgical resident to see (Figure 1). To this end, the Microsoft Kinect sensor version 2 (Microsoft Corporation, USA) was used as a mechanism of touchless interaction – enabling the system to be used in the sterile operating field. Refer to Feng et al. for a detailed description of the system and how it works.

Experiment Design and Procedure

The experimental design is a counterbalanced, within-subject design, with two mentoring approaches: the control is Standard condition, and the intervention is Virtual Pointer condition. In the Standard condition, trainer instruction was conducted as it would be normally, through verbal or hand gestures. In the Virtual Pointer condition, the Virtual Pointer application was used by the trainers as an addition to standard guidance to facilitate instruction.

The trainees worked on four simulated laparoscopic tasks under trainer guidance. The tasks were selected based on a hierarchical task analysis of the laparoscopic cholecystectomy procedure and confirmed by an attending surgeon that they were of similar difficulty levels and required both skills of anatomical structure identification and instrument manipulation. The tasks were performed on a validated laparoscopic training physical model, including (1) mobilizing the cystic duct and the cystic artery, (2) clipping the cystic duct, (3) clipping the cystic artery, and (4) cutting the cystic artery and the cystic duct. Task order and condition were counterbalanced for each trainee yielding a total of 14 runs in the Virtual Pointer condition and 14 runs performed in the Standard condition.

The study was approved by the University of Maryland, Baltimore County institutional review board (IRB) and informed consent was obtained from all participants before their participation. After consent, the trainees and trainers completed a demographics questionnaire, which included information on their surgical experience and familiarity with the Kinect system. After each task, the trainees and trainers completed a cognitive load questionnaire. The study was video recorded and the operative field was screen recorded.
Participants

Participants were recruited from the Department of General Surgery, Anne Arundel Medical Center, Annapolis, Maryland - 7 surgical trainees (all male), including 1 surgical fellow, 1 research fellow, and 4 surgical residents (2 PGY-1 and 2 PGY-2) were recruited. One attending surgeon and one surgical fellow were recruited as the trainers. The attending surgeon guided the surgical fellow in performing the tasks and the surgical fellow guided the rest of the trainees. None of the participants had any previous interactions with the Virtual Pointer before the experiment.

Data Collection

Subjective Measure of Cognitive Load. Subjective measures are considered an unobtrusive, and inexpensive method of assessing workload and are commonly employed to assess surgeons' cognitive load in training scenarios. Commonly employed cognitive load questionnaires for training are the Paas scale, a short, single 9-point Likert scale; and the NASA-TLX, a considerably longer, multidimensional scale. Given that the NASA-TLX requires more time to complete and the fact that it aims to assess overall workload, i.e., not only cognitive load, we chose the Paas scale for our study. The trainees completed it with reference to the mental effort they invested in understanding the instructions, from 1 – very, very low mental effort was invested; to 9 – very, very high mental effort was invested. The trainee completed the questionnaire after performing each task. The Paas scale was also used by the trainers to examine their cognitive load with reference to the mental effort they invested in giving the instructions.

Objective Measure of Cognitive Load. We also included an objective, continuous physiological measure of cognitive load – EDA (electrodermal activity). This continuous measure allows us to assess temporal, dynamic aspects of cognitive load. Physiological measures have a high degree of sensitivity; allowing them to be used for measuring changes in cognitive load. The Empatica E4 wristband was used to record the trainees’ EDA used as the objective measure of cognitive load. The Empatica E4 wristband (Empatica Inc., US) was worn on the trainees’ non-dominant wrist. According to the E4 user manual, the nondominant wrist is recommended to minimize motion artifacts. At the start and end of each task, an event was tagged on the wristband through its physical button. These tagged events were to be used in the data analysis process for time synchronization. The data was recorded directly onto the E4’s flash memory, i.e., wirelessly. After recording each trainee, the E4 was connected via USB to a PC, and Empatica Manager was used to transfer the data to Empatica’s secure cloud platform, Empatica Connect. Each trainee’s data was then downloaded and renamed with her/his anonymized participant ID. The start and end time of each task was tagged directly on the data using the physical button on the wristband.

Objective Measure of Trainee’s Response to Trainer’s Instruction. Video recordings were annotated to determine how often the trainee stopped performing the task to listen to instruction (i.e., single-task instances) and how often they continued performing the task while also listening to instruction (i.e., dual-task instances). All non-study-related events that occurred between tasks, such as waiting for equipment or fixing of the model were not considered in the analysis. In both mentoring conditions, only the trainees’ responses for each instruction given by the trainer were considered. In the Virtual Pointer condition, instructions using the Virtual Pointer, such as moving it to indicate a structure or using it to facilitate verbal instruction were annotated. In the Standard condition, verbal instructions such as “over here” and “move up higher” were also annotated.
**Objective Measure of Number of Turns and Turns Length.** To examine communication efficiency between the trainer and trainee, we analyzed the turn taking structure of communication. As communication becomes more efficient, turn-taking increases and turns become shorter \(^27\). The video recordings were annotated to count the number of turns the trainee and trainer took throughout each task. Both verbalizing and taking an action (e.g. the trainer giving an instruction and the trainee performing a laparoscopic action) were counted as turns. The length (i.e. amount of time) of each turn was also measured. As the number of turns throughout each task is associated with the length of the task, the fraction of number of turns to the sum of all turns’ length was computed as a normalized measure for each task. The analysis was performed on the normalized number of turns and length of turns.

**Data Analysis**

**EDA Data Preprocessing.** Observer XT version 14.1 (Noldus, Netherland) was used to synchronize the EDA data with the video recordings. Each trainee’s EDA data was normalized to a range between 0 and 1 as is standard practice \(^45\). For each specific comparison, normalization was performed across tasks which enabled us to compare subsequent experiences of each mentoring condition for each trainee. We focused our analysis of the EDA data around the points of instruction provided by the trainer in both mentoring conditions. We identified the times that the trainer used the pointing and drawing mode to show a location on the laparoscopic display and the corresponding points of instruction in the Standard condition. A verbal instruction in the Standard condition was considered corresponding to a Virtual Pointer instruction if it was given to show the trainee a location (e.g. “right in the middle”). The instructions were considered as stimuli in the EDA signal. The EDA signal has two components: the slowly changing skin conductance level (SCL) component, and the rapidly changing skin conductance response (SCR) component. The SCR occurs as a peak in the signal, generally as a response to stimulus between 1 to 4 seconds after stimulus presentation \(^46\). Therefore, we extracted the EDA data for the first 5 seconds after the start of each instruction, Peak amplitude of the SCR has been used as an indicator of cognitive load \(^47\). The EDA signal was decomposed into SCL and SCR by performing continuous decomposition analysis using the Matlab-based software, Ledalab \(^48\). The amplitudes of the SCRs for the five-second intervals were computed and used in the statistical analysis.

**Statistical Analysis.** Statistical analysis was performed using a linear mixed model to compare the trainees’ cognitive load between the two conditions for both subjective and objective measures. Because our intention was to focus on the effect of the Virtual Pointer on cognitive load over time, we modeled the mentoring conditions (Virtual Pointer or Standard) and the task order as fixed factors. Due to the similar level of difficulty of the tasks, we considered the task as a random factor. The trainees were also considered a random factor. All statistical analysis was performed using R version 3.2.0 (R foundation for Statistical Computing, Austria). The results are shown as mean and standard error of the mean. A p-value of less than 0.05 was considered statistically significant.

**Results**

**Trainees’ subjective cognitive load scores**

The trainees’ average self-reported cognitive load scores were a bit higher in the Virtual Pointer condition (M=4.28, SD=1.72) compared to the Standard condition (M=3.78, SD=1.67) indicating that the trainees did not experience a significant increase of cognitive load when instructed using the Virtual Pointer (p=0.224) (Figure 2.a). We then investigated whether the trainees’ self-reported cognitive load scores decreased over time, meaning the second experience of using the Virtual Pointer perceived as less cognitively demanding in comparison to a second experience with the Standard method. We analyzed the differences in cognitive load scores between the first and second experience of each condition (Figure 2.b). Although the average cognitive load scores decreased by the second experience of each condition, the decrease within each condition was not significant. However, because the scores did decrease, we also analyzed the decrease between the second experience of each condition. Somewhat surprisingly, these results were not significant either; suggesting that the trainees’ perception of cognitive demand while using a Virtual Pointer is not significantly different from the Standard instruction.

Given that the trainers’ cognitive load is also likely to be impacted by using the Virtual Pointer as an instruction mechanism, it was important to assess whether different levels of cognitive effort are required to provide the instruction. We compared the trainers’ average self-reported cognitive load scores across all tasks using the Virtual Pointer with all tasks in the Standard condition (refer to Figure 2.c). The results indicate that the trainers’ cognitive load actually decreased significantly with the Virtual Pointer (M=3.42, SD=1.45) in comparison to the Standard condition (M=5.64, SD=2.06) (p<0.001). This result suggests that the Virtual Pointer is more beneficial to use as an instructional aid than the traditional instruction method.
EDA Data and SCR Amplitudes

We also examined the trainees’ EDA as it is an involuntary, continuous objective measure of cognitive load. Statistical analysis was conducted using the SCR amplitudes for the first 5 seconds after the start of an instruction with the Virtual Pointer and equal instructions in the Standard condition. A total of 62 Virtual Pointer instructions and 35 equal instructions in the Standard condition were identified. Overall, no significant difference was found in SCR amplitudes between the Virtual Pointer (M=0.031, SD=0.057) and the Standard condition (M=0.025, SD=0.033) (p=0.576) (Figure 3.a). This result confirms the finding of the subjective cognitive load measure.

To have a better understanding of how additional familiarity with the system may impact cognitive load, we also looked at the SCR amplitudes between the first and second experience of each condition as illustrated in Figure 3.b For the Standard condition, similar to the subjective results, the decrease in SCR amplitude was not significant (first experience: M=0.030, SD=0.059; second experience: M=0.026, SD=0.021) (p=0.606). However, for the Virtual Pointer condition, there was a slightly significant decrease in SCR amplitudes from the first experience (M=0.037, SD=0.071) to the second experience (M=0.024, SD=0.034) (p=0.047). These results suggest that although there is an initial slight cognitive load on the trainees with the first experience with the Virtual Pointer system, this load appears to diminish by the second experience as the trainees become accustomed to the system. In fact, after the second experience of instruction with the Virtual Pointer, cognitive load in the Virtual Pointer condition is not significantly different from the traditional mentoring method of instruction.

Single- and Dual-task Instances

To investigate whether cognitive load may be attributed to how the trainees attended to the tasks at a fine-grained, instruction-by-instruction level, the occurrence of single- and dual-task instances was analyzed for each condition. Overall the percentage of single-task instances was higher than dual-task instances in both the Virtual Pointer condition (60.65%) and Standard condition (77.8%). To compare the two mentoring conditions, the percentage of single- and dual-task instances in each of the four laparoscopic tasks was normalized to a range between 0 and 1. Overall, there was no significant difference in the percentage of dual-task instances between the two mentoring conditions (p=0.146) even though the mean was higher in the Virtual Pointer condition (Virtual Pointer M=0.392, SD=0.152; Standard M=0.238, SD=0.373). To further examine the dual-task instances, the percentage of dual-task instances within and between the first and second experience of both conditions was analyzed (refer to Figure 4.a). The results show not only that the increase in dual-task instances within the Virtual Pointer condition was significantly higher in the second experience (p=0.01), but also that dual-task instances completely disappeared in the second experience of the Standard condition (p=0.004). Thus, the differences between the second experience of each condition are also significant (p<0.001) (Virtual Pointer: M=0.476, SD=0.115; Standard M=0, SD=0). These overwhelming differences in dual-task instances in the second experience of the Virtual Pointer suggest that as the trainees became more accustomed to the system, they were able to perform more of the laparoscopic task while watching and listening to the trainer’s instructions without negatively impacting their cognitive load.

---

**Figure 2.** (a) Overall comparison of cognitive load scores of trainees between the Virtual Pointer and Standard conditions. (b) Comparison of cognitive load scores between the first and second experiences of each mentoring condition. (c) Comparison of trainers’ cognitive load score between the Virtual Pointer and Standard conditions.
Figure 3. (a) Comparison of the SCR amplitudes between the Virtual Pointer and Standard condition. (b) Comparison of the SCR amplitudes between the first and second experiences of each mentoring condition.

Figure 4. (a) Comparison of dual-task percentage between the 1st & 2nd experiences of conditions. (b) Comparison of the turns length mean between conditions. (c) Comparison of the number of turns mean between conditions.

Turn Taking Results - Number of Turns and Turns Length Mean

To investigate whether the higher number of instances of dual-tasks while experiencing the same level of cognitive load in the Virtual Pointer condition is due to more efficient communication, the turn-taking structure, specifically the normalized number of turns (Figure 4.c) and turns length (Figure 4.b), was analyzed. Overall, there was a significant difference ($p=0.01$) in the turns’ length between the Virtual Pointer condition ($M=1.975, SD=0.507$) and Standard condition ($M=2.535, SD=0.889$). In addition, significantly higher normalized number of turns ($p=0.01$) were found in the Virtual Pointer condition ($M=0.623, SD=0.142$) compared to the Standard condition ($M=0.487, SD=0.138$). The shorter turns’ lengths and higher number of turns in the Virtual Pointer condition supports our assumption that the Virtual Pointer improves the communication efficiency, thus allowing for further cognitive processing. Consequently, trainees are able to perform dual-tasks while experiencing no additive cognitive load.

Discussion and Conclusion

The goal of this study was to evaluate a Virtual Pointer system in terms of how the added visual information affects trainee’s cognitive load, i.e. the effort needed for trainees to process instruction and task related information. Our hypothesis was that although the Virtual Pointer would add more visual information, it could improve communication efficiency between trainer and trainee, thereby allowing more space for cognitive processing.

Overall, we found no additional cognitive load was associated with the Virtual Pointer after trainees became accustomed to the system despite the indication of prior work. There was a slight non-significant increase to the trainees’ cognitive load in the first experience of the Virtual Pointer condition compared to Standard condition. However, despite the addition of a visual source of information, i.e., a component seen to have the potential of causing increased cognitive load on trainees in a CSCL environment, it in fact does not impose more cognitive load on...
trainees. In fact, once the trainees gained additional experience with the Virtual Pointer-based instruction, it was less cognitively demanding 20,21,50. This result is also in line with our previous findings that the benefits of a virtual pointer in improving the trainees’ performance become evident after initial knowledge is gained 37.

Our further analysis shows the Virtual Pointer could improve communication efficiency between trainer and trainee. Trainees were able to perform more dual tasks – the laparoscopic task itself and watching/listening to the trainer’s instruction – with the Virtual Pointer without being cognitively overloaded. Moreover, an increase in dual task instances in the Virtual Pointer condition did not lead to higher cognitive load. This is especially interesting considering previous research has demonstrated that performing two tasks simultaneously is more cognitively demanding 44. Our findings that increased dual-tasks did not increase cognitive load suggests that the Virtual Pointer actually improves the communication efficiency between the trainee and the trainer, allowing the trainees to have further cognitive processing as suggested by Lim 18. According to the modality effect in Cognitive Load Theory, receiving both verbal and visual instruction leaves more cognitive capacity for trainees to perform more cognitive processing. Thus, the combined visual and audio instruction provided with the Virtual Pointer may open more cognitive processing space enabling the trainees to listen to instruction and perform the surgical task simultaneously. This contrasts with the Standard condition in which less efficient communication, i.e only audio guided instruction, requires greater mental effort to visualize the instruction and perform the laparoscopic task simultaneously. Being able to perform two tasks simultaneously – the laparoscopic task in addition to watching and listening to the trainer’s instruction - may contribute to better comprehension and thus better performance as was indicated in 37 due to benefits of embodied learning 51. The embodied learning theory states that learning and consolidating mastery occurs better by physically practicing the task than by mentally simulating it. Therefore, the better performance of trainees while using a Virtual Pointer 37 might be the result of embodied cognition which is gained through performing and listening to the instruction at the same time.

Limitations

This study was conducted in a simulated training environment. However, the Virtual Pointer is intended for instructing trainees while they are performing laparoscopic surgical tasks on real patients in the operating room - a more complex, stress-inducing, and cognitively demanding condition. In the operating room, even a slightly higher cognitive load may result in poorer performance which could be detrimental to real patients. Thus, further studies need to be conducted to evaluate the use of the Virtual Pointer and possible unintended consequences to workflow in a real life operating room environment. In addition, while EDA has been recognized as the most precise physiological signal for measuring cognitive load 52, it is still affected by other factors, such as activity and movement. Although we removed the segments of EDA data identified to be influenced by factors other than cognitive load, there is still a risk that some of the observed changes in EDA may have other underlying origins not associated with cognitive load. One concern is distinguishing stress from cognitive load in EDA-based measurement methods. However, results from 49 on discriminating between stress and cognitive load demonstrate that SCR interval rate is a predictor of stress, while the number of SCRs present, used as our measure, is a predictor of cognitive load. Additionally, use of the non-dominant wrist for collecting EDA data, although recommended by the Empatica E4 user manual, might miss some relevant information related to cognitive load from the dominant hand.

Acknowledgements

The authors gratefully acknowledge the Anne Arundel Medical Center for the use of equipment and space in the Simulation to Advanced Innovation and Learning (SAIL) Center, and would like to thank Ms. Katie Li and Ms. Jacqueline Mun for their support in data collection and analysis, as well as the participants who devoted their time to this study. This work was supported by NSF Grants IIS #1422671 and #1552837.

References

Using FHIR to Construct a Corpus of Clinical Questions Annotated with Logical Forms and Answers

Sarvesh Soni, MS1, Meghana Gudala, MBBS1, Daisy Zhe Wang, PhD2, Kirk Roberts, PhD1
1School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
2Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL

Abstract

This paper describes a novel technique for annotating logical forms and answers for clinical questions by utilizing Fast Healthcare Interoperability Resources (FHIR). Such annotations are widely used in building the semantic parsing models (which aim at understanding the precise meaning of natural language questions by converting them to machine-understandable logical forms). These systems focus on reducing the time it takes for a user to get to information present in electronic health records (EHRs). Directly annotating questions with logical forms is a challenging task and involves a time-consuming step of concept normalization annotation. We aim to automate this step using the normalized codes present in a FHIR resource. Using the proposed approach, two annotators curated an annotated dataset of 1000 questions in less than 1 week. To assess the quality of these annotations, we trained a semantic parsing model which achieved an accuracy of 94.2% on this corpus.

Introduction

Fast Healthcare Interoperability Resources (FHIR)1 is an emerging standard developed by Health Level Seven International for storing and sharing information between different electronic health systems. Because of its RESTful architecture (adherence to REpresentational State Transfer style), it can be effectively used to view the data in different settings such as mobile and web applications. However, accessing some required information in electronic health records (EHRs) is still a cumbersome task2. Efficient query mechanisms such as question answering (QA) from medical records have the potential to reduce the delay between stored information and its prospective users.

EHRs contain information in both structured and unstructured format. Information extraction from unstructured data present in EHR (such as clinical notes and diagnostic reports) is well-researched and holds a promise to convert the free text data into structured format3,4. Efficient techniques to surface information from such structured EHR data have the potential to make the overall information access faster and reduce burdens related to system complexity.

Systems for QA from structured databases rely on unambiguous interpretation of natural language questions, often represented as machine-understandable logical forms. This process of converting natural language questions to their logical representations is called semantic parsing. Training semantic parsers usually requires a set of questions along with their annotated logical forms and/or answers5,6. Such corpora are widely available for the general domain but there is a scarcity of such datasets for EHR QA, mainly because of privacy issues and the complexity of EHR data.

Though annotating the questions with answers seems to be an easier task, one of the main difficulties involved in training a semantic parser using such dataset is ambiguity related to the realization of logical forms. Multiple logical forms can produce the correct answer but it becomes challenging for the semantic parser to separate the correct logical forms from the incorrect ones in such a scenario. Moreover, constructing a dataset of question-logical form pairs require annotating concept normalizations where medical concepts are recognized and mapped to a standard ontology. This is one of the hardest and time-consuming parts of creating a dataset for semantic parsing7.

Concept normalization alone is a challenging task as the same basic concept may legitimately map to multiple clinical codes and/or be stored in separate FHIR resources, even with the interoperability standards in place. For instance, the information that a patient has abdominal tenderness can be stored using FHIR Observation resource in at least the following ways8:

i. code: C4321457 (Examination), value: Abdomen tender
ii. code: C0562238 (Examination of abdomen), value: Tenderness
iii. code: C0232498 (Abdominal tenderness), value: found or true
iv. code: C0232498 (Abdominal tenderness), value: no value

1207
Moreover, a Condition resource can also be used to store this information if the symptoms are lasting\(^9\).

One of the main reasons behind such disparities is the use of different practices to store information at different organizations, which makes the process of normalizing concepts harder in a real setting. While there exist automated methods for concept normalization\(^9\), they are not tailored to the practices of a specific organization. Hence, it is important to learn these conventions at the organizational level where the QA system will be deployed.

In this paper, we propose a novel approach to construct a sizeable corpus of question-logical form-answer triplets using a FHIR server in a relatively short amount of time. Precisely, the list of available FHIR resources for a patient is shown to the annotators who, then, create a question, select the correct answer, and construct the corresponding logical form. We make use of the FHIR resource answers to automatically annotate concept normalizations for highlighted question phrases, which reduces the overall annotation time. Further, we train a semantic parser over the constructed dataset to learn the underlying logical forms. We also implement a concept normalizer tailored to the annotated concepts present in our corpus. Because of the shorter time and lesser prerequisite knowledge requirements for such corpus construction, the approach can be scaled across the organizations.

**Background**

We divide the related work into the following two sections based on the type of data used by medical QA systems.

1. **Unstructured Data**

Numerous studies in medical QA focused on the unstructured data sources such as biomedical literature, health-related social media data, and free text EHR data\(^{1,12}\). Several studies aimed at extracting relevant documents from the biomedical literature\(^{13–16}\). The BioASQ challenge\(^{17}\) introduced a decent sized dataset for machine comprehension (MC), i.e., answering a question on the basis of a given free text, from biomedical scientific articles. Similarly, a small MC dataset was released as part of the QA4MRE task\(^{18}\). Further, a recent study created large scale MC dataset using PubMed\(^{19}\). Another work constructed a small dataset using the resolved question-answer pairs from community-based health forums\(^{20}\). The MedQA task\(^{21}\) involved a large scale reading comprehension dataset which was built using data from certification exam for medical practitioners. Using the same certification exam dataset, a study proposed a framework for automatically generating question-answer pairs\(^{22}\). Further, with the MedicalQA task\(^{23}\), a dataset was developed utilizing various components of EHRs such as history of present illness. Raghavan et al.\(^{24}\) defined a process of annotating question-answer pairs using the EHR clinical notes.

Moreover, several efforts aimed at structuring the data present in unstructured form. Goodwin and Harabagiu\(^{25}\) proposed an approach for converting MIMIC-III medical records to knowledge graph for QA. Another study by Ayalew et al.\(^{26}\) created an ontology for the dataset of online frequently asked questions to assist in QA.

The majority of the datasets proposed for medical QA from unstructured data are based on biomedical literature whereas, only recently, the unstructured EHR data is gaining attention. Nevertheless, many studies are directed toward information extraction from free text present in EHRs\(^5\). Our aim, however, is to construct a corpus using the structured EHR data.

2. **Structured Data**

Very few works concentrated on QA from the structured biomedical data. Asiaee et al.\(^{27}\) and Amith\(^{28}\) worked on QA over a medical ontology. Pampari et al.\(^{29}\) constructed question-logical form and question-answer pairs using a template-based approach. QA dataset generation using templates limits the variety of questions to a certain extent and are not representative of the real world queries. Our previous work included annotating a set of EHR questions with their corresponding logical forms\(^7\). As stated earlier, logical form annotations are time-consuming to annotate, especially because of difficulties involved in annotating the concept normalizations. In this paper, we aim to quicken the process of concept annotations utilizing FHIR. To our knowledge, this is the first work to construct a dataset for EHR QA using FHIR resources.

**Materials and Methods**

The following sections describe the methods used for constructing and evaluating the dataset. A graphical representation of our methods is shown in Figure 1.

1. **Dataset Construction**

We set up a local FHIR server which serves as an underlying data source for constructing the dataset. An annotation tool is implemented to present the information contained in the FHIR server to the annotator. The annotators reviewed
the available patient resources and constructed questions along with their corresponding simple logical forms. An institution-specific concept normalizer is trained based on these annotations. We further processed the constructed questions to be used in training a semantic parsing system. Each of these tasks is explained further in the following sub-sections:

1.1. FHIR Server

We deployed a FHIR server locally using an existing open source implementation by the MITRE Corporation for DSTU2 version (v1.0.2)\(^*\). To generate the data for this local server, we used Synthea, a tool for producing realistic health care records\(^3\). The FHIR resources generated by Synthea are added to the local FHIR server using the REST APIs provided by the server. We chose to use the local FHIR server instead of any online available sandbox servers to maintain stability and consistency.

1.2. Annotation Tool

We implemented a web application to facilitate the browsing of FHIR resources present in the local server for an efficient annotation process. The tool displays all the available patients along with their resources from the server. The three main components of the tool interface allow, for a selected patient, viewing all the resources, viewing all the resources of a selected resource type, and entering the constructed question with its corresponding logical form. The user interface for the annotation tool is shown in Figure 2. The main components of the tool are explained below. Indices in the parentheses, (#), point to the corresponding elements in Figure 2.

I. Patient Details

This view gives an overview of all the FHIR resources for a selected patient. A list of available patients along with a count of questions constructed for them can be seen using the dropdown (1). The patient can be changed using this dropdown or by refreshing the browser. The gender and date of birth information are shown below the dropdown at (2) for a selected patient.

The table in this component shows the information common to all the FHIR resources, namely, resource type, name, ID, and time. All the entries in this table can be capped, filtered, and/or sorted using the corresponding functionalities at (3), (4), and (5). Such flexibility in viewing the resources facilitates the overview of patient history.

II. Selected FHIR Resource

This part enables the annotator to view more granular details for the resources of a selected type. All the available resource types for the selected patient can be seen in the dropdown (6).

\(^*\) https://github.com/mitre/fhir-server
The table under this section has similar functionalities in terms of filtering and viewing the results. In addition to these, the table allows selecting one or more resources from the list using checkboxes (7). This selection serves as the answer to a constructed question.

III. Annotations

Above this section at (8) is an ongoing count of the total number of questions the annotator has constructed so far. This count takes into account the questions constructed for all the patients to show the overall annotation progress. The question text box in this component allows the annotator to input their constructed question. The concept boundaries can be highlighted using a set of opening and closed square brackets. The marked text is automatically normalized to the medical concept present in the selected answer. The text box for simple logical form is used for entering the logical form for the constructed question. There is another text box for entering the comments in regard to the annotation, if there are any. The submit button is used to save the annotation to the corpus.

The tool is equipped with certain validation checks which ensure that a correct annotation is saved. Specifically, the tool checks for the following:

- a. At least one FHIR resource is selected as an answer
- b. The question contains a highlighted concept
- c. The simple logical form is valid (check for balanced parentheses and valid logical predicates)

1.3. Annotation Process

The description of the annotation tool provides some intuition to the annotation process. In this section, we explain the overall annotation process in detail.
The annotator starts with a selected patient and views all the FHIR resources for them in the Patient Details component. The annotator decides to ask questions based on the available resources of a patient. After getting a quick overview of the patient, the annotator moves to the Selected FHIR Resource component and selects a resource type to construct some question. The annotator enters a question in the Question input and selects the corresponding FHIR resource(s) containing the answer. Finally, the user enters a corresponding simple logical form and hits submit. We follow the logical form structure as described in our previous work on annotating the EHR questions2. After the submission of an annotation, the highlighted concept in the question text is automatically normalized to the medical concept of the selected FHIR resource(s). The annotation is saved to the corpus with this normalized concept. For instance:

**Answer Resource:** Condition resource Rupture of appendix with concept code C0267628

**Question:** When was the [appendix ruptured]?

**Normalized Question:** When was the nn:concept(C0267628)?

**Simple Logical Form:** time(latest(lambda(concept)))

**Logical Form:** time(latest(λx.has_concept(x, C0267628)))

Again, only the Question and Simple Logical Form are typed by the annotator. The Answer Resource comes from the checkbox (7) and the Normalized Question is computed automatically.

### 1.4. Post Processing

After the annotation process is completed, the constructed corpus is further processed to add more annotations. Particularly, we employ rule-based approaches to identify the person references, temporal references, and measurements. These additional annotations are important to separate the task of semantic parsing from concept normalization4. The following question exemplifies such normalization categories:

**Question:** Did her hemoglobin A1c exceed 6% in the past 3 years?

**Normalized:** Did pos:person nn:concept exceed measurement(‘6%’) in the temporal_ref(‘past 3 years’)?

These additional references are easier to annotate automatically because a small number of patterns can cover the majority of cases. For instance, the person references can be found in the question text using a set of patterns such as the patient, she, he, her, and him. The temporal references are identified using the regular expression patterns, e.g. yesterday, past n years, and last n days. Similarly, the measurements are extracted by employing patterns like n units where units can be C, F, lb and so on. These rules for identifying the additional concepts are motivated by our previous work.

### 2. Evaluation

For assessing the quality of annotations using our tool, we employ a concept normalization and a semantic parsing model. Each of these models is separately explained in this section.

#### 2.1. Concept Normalization

For assessing the effectiveness of learning an institution-specific concept normalizer, we use simple deep learning models based on convolutional neural network (CNN) and recurrent neural network (RNN) architectures. Specifically, we follow the technique inspired by LimSPatham and Collier51, which make use of the publicly available pre-trained embeddings to normalize medical concepts from health-related social media messages. We randomly initialized the embeddings for the words present in questions and concepts.

To train these models, we input the highlighted concept phrase in the question along with the automatically annotated concept term from FHIR. Using the above example of Rupture of Appendix,

**Highlighted phrase in question:** appendix ruptured

**Annotated concept:** Rupture of Appendix (C0267628)

The concept normalizer learns these mappings from the dataset and aims to translate the medical terms present in natural language text to the standard medical concepts. Note that we do not train a concept boundary classifier, as this is an extremely common clinical natural language processing (NLP) task32–36.
2.2. Semantic Parsing

We train a semantic parser over the question and logical form annotations following a hybrid approach using both rule-based and machine learning-based techniques. Full details of the employed method can be found in our previous work on semantic parsing\(^6\). We update the lexicon, which maps lexical phrases present in questions to the corresponding logical operations, to extend its coverage. Moreover, we use similar rules for candidate generation and same features for the machine learning model as used in our previous approach.

Results

Two annotators (a physician and a biomedical informatics doctoral student) were involved in constructing the corpus of 1000 questions. Each annotator independently created 500 questions following the annotation guidelines presented in the methods section. The individually constructed annotations were cross reviewed by both the annotators after the completion of the first 50 and all 500 annotations to ensure the completeness. We stored the timestamps for each annotation which gave us the ability to analyze our annotation times in detail. We kept track of the review times manually. The first checkpoint of 50 annotations was completed in 2.5 hours by each annotator and was reviewed by both the annotators together in 1 hour. Each annotator took about 24 hours to complete the remaining 450 annotations, which were then separately reviewed by both the annotators in 8 hours. The whole annotation process was completed in less than one calendar week. By contrast, Roberts and Demner-Fushman\(^7\) required multiple weeks to annotate less than half as much data, while their concept normalization choices were likely incompatible with many hospitals as it was not built from a reference FHIR instance.

Figure 3. Frequency distribution of FHIR resource types selected as answer in the constructed corpus.

Figure 4. Frequency distribution of answer types in the annotated corpus. Types (y-axis) denote the final function applied over the answer resource(s).
Table 1. Most frequently co-occurring FHIR resource and answer types along with their corresponding counts. 

<table>
<thead>
<tr>
<th>Resource type</th>
<th>Answer type</th>
<th>Frequency</th>
<th>Example</th>
</tr>
</thead>
</table>
| Observation   | Value       | 92        | Q: What is his t-score?  
|               |             |           | LF: latest(\(\lambda x\).has_concept(x, C1526354)) |
| Observation   | Max/Min     | 60        | Q: What was the highest hemoglobin A1c value in the past 2 years?  
|               |             |           | LF: max(\(\lambda x\).has_concept(x, C0366781) \& time_within(x, 'past 2 years')) |
| Observation   | Is state    | 59        | Q: Has her calcium level ever been less than 9 mg/dl?  
|               |             |           | LF: delta(\(\lambda x\).has_concept(x, C1977516) \& less_than(x, '9 mg/dl')) |
| Condition     | Is state    | 54        | Q: Was her ankle sprain healed?  
|               |             |           | LF: is_healed(latest(\(\lambda x\).has_concept(x, C0160087))) |
| Condition     | Time        | 53        | Q: When did he develop microalbuminuria?  
|               |             |           | LF: time(latest(\(\lambda x\).has_concept(x, C3875084))) |

Table 2. Results of evaluation. RNN – Recurrent Neural Network, CNN – Convolutional Neural Network.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic Parser</td>
<td>94.2%</td>
</tr>
<tr>
<td>Concept Normalizer</td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>76.0%</td>
</tr>
<tr>
<td>CNN</td>
<td>70.0%</td>
</tr>
</tbody>
</table>

We observed a total of 10 unique FHIR resource types which were used as answers in our corpus. The distribution of these resource types is presented in Figure 3. We note that the most number of constructed questions were about Observation resource followed by Condition and Procedure resources.

Based on the annotated logical forms, we enumerated 11 answer types. Figure 4 shows the distribution of these answer type frequencies in our annotated corpus. Each type can be perceived as a logical predicate applied to the FHIR resource(s). For e.g., Time answer type is assigned to questions which return the time of a specific resource. Similarly, Is state type aims to determine the state of some specific resource(s). It should be noted that these answer types are based upon the outermost logical predicates in logical forms.

We analyzed the combinations of FHIR resource and answer types in our dataset to get the impression of most frequent question varieties. We present the top 5 of these combinations with example question and logical forms in Table 1. The most frequent number of questions, specifically 92, were about querying the value of Observation resource, which succeeded by querying the Max/Min and Is state.

The results of training semantic parser and concept normalizer are shown in Table 2. Using rules and features of the best performing model from our previous approach9, we achieved an accuracy of 94.2% with leave-one-out validation. We trained the concept normalizer models based on RNN and CNN for 50 epochs each, as per the evaluation results presented in the original paper91, and achieved an accuracy of 76.0% and 70% respectively.

Discussion

Annotating concept normalizations during the logical form annotation is a challenging and time-consuming task. We aimed at automating this step using the highlighted concept text in question. This intervention reduced a significant amount of time for annotating the logical forms in comparison to our previous approach7. Another advantage of this approach is that the annotated concepts in our corpus are well-aligned with concepts used in the referenced FHIR server. This enabled us to train a concept normalizer tailored to the FHIR server used for annotations.

The result of the semantic parsing evaluation highlights the quality of generated corpus. The accuracy of the semantic parser on our corpus is slightly less than that on the dataset it was originally built on. This difference might have resulted from increased vocabulary and variety of questions in our corpus.

The decent performance of the concept normalizer shows the potential of this annotation approach for concept normalization. In comparison to other methods32-36, we trained on less data. Also, our corpus contains a wider variety of concepts than just disorders or diseases. Other concept normalizers such as DNORM37 could be trained but that was not the focus of this work.
Using the proposed annotation approach, a large annotated dataset of questions can be created comparatively faster. This dataset can then be used for building a semantic parser to support natural language queries.

We used a rule-based approach to identify the temporal expression spans, as this approach was able to capture the variety of expressions in our corpus. More sophisticated temporal information extraction systems can be incorporated in the future.38,39

One of the limitations of our study is that we limit the answer resources selection to a single concept type. In other words, the constructed questions could be about a single type of concept. In the future, we aim to include multiple concept types during answer selection which can result in the creation of more complex questions.

Conclusion

We have described a novel approach for constructing an annotated corpus of EHR questions using FHIR. The process of annotating concept normalizations is automated which significantly reduced the overall annotation time. Using the proposed approach, two annotators created a corpus of 1000 questions in less than one week. We evaluated our constructed dataset by training a semantic parser and a concept normalizer, both of which showed promising results with accuracies of 94.2% and 76% respectively.

Acknowledgments

This work was supported by the U.S. National Library of Medicine, National Institutes of Health (NIH), under award R00LM012104, as well as the Bridges Family Doctoral Fellowship Award.

References

Learning Inter-Sentence, Disorder-Centric, Biomedical Relationships from Medical Literature.

Anton H. van der Vegt, BE, BSc 1Guido Zuccon, PhD 2Bevan Koopman, PhD
1The University of Queensland, St Lucia, Qld, Australia; 2CSIRO, Brisbane, QLD, Australia

Abstract

Relationships between disorders and their associated tests, treatments and symptoms underpin essential information needs of clinicians and can support biomedical knowledge bases, information retrieval and ultimately clinical decision support. These relationships exist in the biomedical literature, however they are not directly available and have to be extracted from the text. Existing, automated biomedical relationship extraction methods tend to be narrow in scope, e.g., protein-protein interactions, and pertain to intra-sentence relationships. The proposed approach targets intra and inter-sentence, disorder-centric relationship extraction. It employs an LSTM deep learning model that utilises a novel, sequential feature set, including medical concept embeddings. The LSTM model outperforms rule based and co-occurrence models by at least +78% in F1 score, suggesting that inter-sentence relationships are an important subset of all disorder-centric relations and that our approach shows promise for inter-sentence relationship extraction in this and possibly other domains.

Introduction

Deriving the relationships between a disorder and its related symptoms, tests and treatments is a critical part of medicine. Consider the relationships that exist between symptoms and their associated disorders in Figure 1. Such disorder-centric relationships are necessary for a myriad of medical tasks and research including automated problem lists1, clinical decision support, medical information retrieval, data mining and knowledge base construction. However, in medical thesauri such as the UMLS and SNOMED CT, such relationships either do not exist, or at least are sporadic. While hand-coding such relationships is prohibitively laborious, it is possible to automatically derive these from free-text using biomedical relationship extraction (BRE) systems. Most systems, however, consider one specific, highly focused relationship set; for example, locations of bacteria2 or protein-protein interactions3. Those that are broader in coverage, and can be used for disorder-centric relationship extraction, exhibit poor performance (in terms of precision and recall).

Figure 1: The title (in bold) and first 2 sentences of Medline article PMID=19707415. Medical concepts have been extracted with MetaMap4 and disorders have been underlined, symptoms have been italicised. Medical terms mapped as both disorders and symptoms are both italicised and underlined.

The old and new therapeutic approaches to the treatment of giardiasis: where are we?

Giardia lamblia is the causative agent of giardiasis, one of the most common parasitic infections of the human intestinal tract. This disease most frequently affects children causing abdominal pain, nausea, vomiting, acute or chronic diarrhea, and malabsorption syndrome.

Figure 1 also highlights another important issue: relationships span across multiple sentences. Without the capability to detect inter-sentence relationships, all of the symptoms (nausea, vomiting, etc.) in the second sentence would not be related to the disorder giardiasis in the title and first sentence. Current state-of-the-art relationship extraction systems such as SemRep5 and PASMED6 only extract intra-sentence relationships. New methods for inter-sentence relationship extraction are needed.

Extracting inter-sentence, disorder-centric relationships involves a number of challenges7. Grammar rules reset between sentences so features, such as shortest dependency path (SDP) analysis, cannot be relied upon. Patterns are harder to identify for inter-sentence, rule-based approaches and the number of possible patterns and/or relationships expands dramatically. To overcome these issues, we recast the inter-sentence relationship extraction task as a sequential labelling problem; a Long Short-Term Memory (LSTM) model is proposed to process a multi-sentence input for a given relationship label occurring across sentence borders. The main contributions of this research are:
1. A deep learning model to perform inter-sentence, disorder-centric biomedical relationship extraction.
2. An empirical evaluation of existing relationship extraction techniques when applied to inter-sentence relationship extraction. In addition, we make available an inter-sentence, disorder-centric, human labelled dataset for training and evaluation purposes, as well as UMLS concept embeddings based on 23 million MEDLINE citations.

**Background and Related Work**

Extracting medical concepts (or entities) from biomedical free-text has had considerable attention. There are well established tools for mapping free-text to UMLS concepts such as MetaMap and QuickUMLS. While these tools identify concepts (including disorders), they do not identify the important relationships that exist between them.

Clinicians tend to ask a common set of medical questions, with the following 4 accounting for 40% of all questions:

1. How should I treat condition $x$?
2. What is the drug of choice to treat condition $x$?
3. What is the cause of condition $x$?
4. What test is indicated in situation $x$?

Based on these findings we focus on extracting relationships between: Disorder and treatments (DT); Disorders and symptoms (DS); Disorders and medical tests (DE); and Disorders and other disorders (DD), as these are often viewed as symptoms for disorders, e.g., hypertension.

We now review previous relationship extraction approaches; these fall into four categories: co-occurrence based, rule based, machine learning based, and more specialised deep learning based.

**Co-occurrence based**: A common approach is to simply infer a relationship between two concepts if they *co-occur* within some window of text (e.g., a sentence). Ding et al. used a variety of window sizes — phrase, sentence, sentence pair and entire abstract — to assess the impact on precision and recall. A window size set at the sentence or abstract level provided the best combination of recall and precision; adjacent sentences produced the worst results. However, interestingly, adjacent sentences nearly doubled the number of distinct relationships found and generated high recall, but also a considerable drop in precision. These insights suggested to us that if the target relationships were much broader than biochemical-noun interactions (relationships) and the selection model more discriminating than the simple co-occur model, then inter-sentence BRE could dramatically improve the recall of biomedical relationships, without a corresponding drop in precision.

**Rule based methods**: incorporate NLP to parse biomedical text and identify specific relationships with either hand-crafted, or automated rules. Rules are created on the basis of syntactic or semantic patterns which describe specific relationship cases; for example, using POS tagging, noun-phrase chunking and dependency parsing to define the rule set. Dependency parse trees enable relationships to be identified where the medical concepts do not appear close together. Other approaches use verbs in the sentence by analysing the phrase-level conjunction to extract relationships.

The main disadvantage of rule-based approaches is that they are usually applied to a narrow set of relationships.

Machine learning methods are commonly employed to perform relationship extraction with supervised classification solutions, such as support vector machines (SVM) and conditional random fields (CRF). Most machine learning methods require careful feature engineering, with a mix of both domain knowledge and NLP experience. Rink et al. trained an SVM on eight relationship types between concepts within a sentence. While 21 features were used, context features proved to be very important and in particular the concept and concept types. In contrast, Uzuner et al. found that lexical features, in particular the tokens occurring between candidate concepts, were most informative in their SVM approach. Quirk et al. employed a distant supervision ML model to extract relationships across sentence...
Table 1: Classification of UMLS concepts into categories for disorder-centric relationship capture.

<table>
<thead>
<tr>
<th>Category</th>
<th>Defining UMLS Semantic Types (abbreviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disorders</td>
<td>Acquired abnormality (acab), anatomical abnormality (anab), cell or molecular dysfunction (comd), congenital abnormality (cgab), disease or syndrome (dsyn), experimental model or disease (emod), injury or poisoning (inpo), mental or behavioral dysfunction (mobd), neoplastic process (neop), pathalogic function (patf)</td>
</tr>
<tr>
<td>Symptoms</td>
<td>Sign or symptom (sosy), Finding (findg)</td>
</tr>
<tr>
<td>Treatments</td>
<td>Health care activity (hlca), therapeutic or preventative procedure (topp), pharmacologic substance (phsu)</td>
</tr>
<tr>
<td>Tests</td>
<td>Lab procedure (lbpr), diagnostic procedure (diap)</td>
</tr>
</tbody>
</table>

boundaries, and this model was used with success for extracting drug-gene interactions from biomedical literature. However, this model did not incorporate UMLS concept-to-concept relationships and the relationships detected were very niche. Like most machine learning approaches, these methods suffer from complex feature engineering.

Deep learning methods have proven effective across a range of biomedical tasks, including relationship extraction\textsuperscript{18}. Deep learning approaches can avoid the laborious feature engineering of the aforementioned machine learning methods. For relationship extraction, a deep autoencoder was used to produce word embeddings of the medical concept word features\textsuperscript{18}. These were then fed into a CRF classifier to identify relationships. Results showed only subtle improvements over the standard feature input set. However, we posit that perhaps using a much broader medical concept vector, across the whole of Medline, could provide greater benefit. Outside the medical domain, long short term memory (LSTM) recurrent neural network (RNN) models have proved highly effective on other tasks\textsuperscript{19}. This approach incorporated shortest dependency paths (SDP) and linguistic information into a multi-channel RNN. Each SDP for each sentence is sequenced from either direction of the SDP using LSTM units and this is done for 4 different channels of information (word representation, POS tag, grammatical relationship and WordNet hypernyms). It was insightful to note that the word embedding had by far and away the greatest impact on performance, and each other channel added at most 1% to the score or combined, only 1.6%. Li et al.\textsuperscript{20} used a similar, bi-directional LSTM model, in the biomedical domain, for the extraction of adverse drug events and bacteria biotopes. Their model also employed SDP information as well as word, POS and character embeddings of the words in the sentence. Gupta et al.\textsuperscript{21} also made use of SDPs, however within a deep learning model incorporating bidirectional RNNs to identify relationships across sentences in biomedical text. This model evaluated with the Bacteria Biotope relationship extraction task, showed very promising performance, however the task incorporated very specific relationships (habitats of bacteria), and unlike our proposed approach, medical text rather than medical concepts were the source of relationships. This is an important distinction because concept space disrupts the natural grammar and syntax of sentences, needed for SDP analysis. The use of SDP within LSTM deep learning models is popular, however for inter-sentence relationship extraction, it is not really possible\textsuperscript{22}. Kim et al.\textsuperscript{7} raised this as a general challenge for participants in the BioNLP shared task in 2009. For this reason, we investigate alternative designs for using the LSTM RNNs to employ for inter-sentence, UMLS concept relationship extraction.

Proposed Disorder-Centric Relation Extraction Model

Typically, deep learning relationship extraction models operate on an intra-sentence basis and therefore the input sequence is limited to, at-most, a sentence of words and often much less if the two entities are closer together. The objective of our approach is to extract relationships within 1 to N sentences, resulting in a significantly larger potential sequence of words between the related entities. To reduce the complexity of this problem, we decided to divide the relationship extraction task into two components: relationship detection and relationship classification. In this way, the deep learning model is employed to detect the presence of a relationship between a disorder and one of treatment, test or symptom concepts, i.e., a binary classification task.

Relationships classification is then done on the basis of the concept types involved in the relationship; for example, if a relationship exists between a symptom and a disorder concept, it is assumed to be a DS relationship. The concept types are based on the UMLS semantic types (semType). SemTypes provide a broad, yet consistent, classification of every UMLS concept. Table 1 defines the semType sets used for classification of concepts into the four types.

Disorders are derived directly from the Disorders semGroup excluding symptoms and findings, which are used in the Symptoms concept category. Tests and treatments semTypes are selected using the semMed DB\textsuperscript{14} predication table. For test semTypes, we counted the predications found for each semTypes involved in a diagnoses relationship and
selected the top 2. For treatments we counted the predictions for semTypes involved in a TREATS relationship and selected the top 3 excluding the Amino Acid, Peptide or Protein semType.

Figure 2 describes the information flow and network architecture of the proposed disorder-centric, relationship detection model. Unlike Xu et al. and Mehryary et al., the sequence components in our model are not divided into feature channels, such as POS, words or grammatical relations. Instead, we use UMLS concepts as the basis of each sequence line and can therefore capitalize on UMLS concept embeddings that span the Medline data set. Concepts have a number of advantages over words for use in embeddings. First, the same medical entity expressed in different ways is aggregated into a single concept making for a denser set of relationships between entities. Second, n-grams are also encapsulated within a single concept — especially important in biomedicine where multi-term entities are common and used as a means of distinguishing them; e.g., over 100 are disorders entitled “[something] hypertrophy” (e.g., pseudo hypertrophy). We posit that embeddings of real biomedical entities (rather than the words ‘pseudo’ and ‘hypertrophy’) are more discriminative. The individual elements of Figure 2 are detailed next.

Input Sentences consist of zero or more title sentences, a target sentence and the sentence before the target (if it exists), called the pre-sentence. The target sentence must contain at least the related concept. The disorder concept may exist in any of the sentences.

CUI-Word Mapping. The input sentences are converted to UMLS concepts using MetaMap (with word sense disambiguation) to extract the best match candidates from each sentence, as well as the POS of each word, and punctuation. Sentences are then converted into a CUI-Word Mapping format, which expresses each word or punctuation mark in the sentence as either part of a CUI, together with its semType, or a word, together with its POS.

CUI Line Sequence Vectors (CLSV) are then generated. The lines are constructed starting from the beginning of the title sentence and moving through to the pre-sentence, finishing at the end of the target sentence. Each CUI that is found along the way, forms a CLSV. A CLSV consists of the words leading up to a single CUI and the CUI itself. Therefore, the sequence will consist of \( n \) lines where there are \( n \) CUIs found across the input sentences. Each CLSV is a concatenation of feature vectors which include:
1. **CUI2VEC embedding for the CUI in the CLSV:** Two CUI embeddings are tested. The first is the 500 dimensional CUI embedding recently generated by Beam et al.\textsuperscript{23} containing 108,477 CUIs, based on 20M clinical notes and 1.7M full text biomedical journal articles. We developed the second using the DL4J word2vec model, which uses Skipgram, CBOW or DBOW feature extraction. The best mapping CUI candidates were extracted from over 23M Medline citations by parsing the 2015 MetaMapped Medline collection. These CUIs were then input to the word2vec model and by applying a window size of 10 and a minimum word frequency of 1 a set of 440-dimensional embeddings were generated for 558,764 CUIs. (500 or 440 dimensions)

2. **Location:** A one-hot vector representation of the location of the CUI in this CLSV. The locations are within either the title, pre-sentence or target sentence. (3 dimensions.)

3. **COSine similarity:** Calculated using the CUI embedding vectors above, the cosine similarities between the CUI in the CLSV and the assessed (i) diagnosis CUI and; (ii) related CUI. Note, the CUIs assessed for relationship presence will appear in at least one CLSV each across the sequence of CLSVs. (2 dimensions.)

4. **SemType:** A one-hot vector representation of the semType of the CUI in this CLSV. (135 dimensions.)

5. **POS tag of up to N words preceding the CUI:** A one-hot representation of the 11 POS tags plus a separate tag for each of full stops, commas, other punctuation and invalid words. In addition a present/not present flag dimension was used to identify all zero POS embeddings for no-word situations. This occurs frequently where less than N words or punctuation marks appear before the CUI. (N x 16 dimensions)

**RNN LSTM model:** An LSTM unit was applied to each sequence line; i.e., each CLSV. For the LSTM, the ‘no peep-hole’ variant\textsuperscript{24} was instanced through DL4J. Standard settings include weight initialization, \textit{tanh} activation function for the LSTM layer and a cross-entropy loss function combined with SoftMax activation for the RNN output layer\textsuperscript{25}.

**Deep Learning Model Tuning & Testing:** During this testing phase a number of model and feature set changes were trialled in order to develop the most effective overall disorder-centric relationship detection system. Important model/feature combinations are reported later in the results section (Table 4). The layer size was set to 300 throughout; batch size to 32 and learning rate to 0.005. From Table 4, the extra LSTM layer (Id=S2) was added in series immediately after the first layer. The dense layer (Id=S3) is a fully connected feed forward layer, connected to the output of the second LSTM layer and feeds into the final RNN output layer.

**Training and Evaluation Methodology**

A **label data set** for training the deep learning model and evaluating the BRE systems was developed. Existing label collections were too specific to utilise; for example, labels derived for protein-protein interactions\textsuperscript{3} or adverse drug events\textsuperscript{26}. The closest suitable label data set, the i2b2 challenge\textsuperscript{4}, was developed for clinical text, not biomedical literature, which is very different, rendering it unsuitable. The high cost of constructing a \textit{gold} label collection — one constructed from expert annotations, of sufficient size — made the gold option not possible. However, based on the reduced requirement for relationship identification, rather than classification, non-expert labellers (via crowdsourcing, with appropriate quality controls) were employed to detect pre-identified concepts within 754 Medline citations. The final label values were assigned through a voting system across multiple labellers (n=5) and multiple tests of the same relationship, found in different parts of the text, with a positive label awarded when the proportion of actual to possible votes for the relationship exceeded a threshold of 0.4. Details of this \textit{silver}, disorder-centric relationship, label set are provided in Table 2. Note the collection is heavily weighted towards negative labels, which is as expected, because across one or more sentence boundaries, the majority of concepts are not directly related.

**Table 2:** Final label counts for the disorder-centric silver label set. Numbers summed by relationship type. DS=Disorder-Symptom DE=Disorder-Test, DT=Disorder-Treatment, DD=Disorder-Disorder.

<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>DS</th>
<th>DE</th>
<th>DT</th>
<th>DD</th>
<th>Totals (% of total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative label</td>
<td>1,927</td>
<td>1,080</td>
<td>1,959</td>
<td>4,064</td>
<td>9,030 (70.6%)</td>
</tr>
<tr>
<td>Positive label</td>
<td>559</td>
<td>658</td>
<td>1,194</td>
<td>1,359</td>
<td>3,770 (29.4%)</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>2,486</td>
<td>1,738</td>
<td>3,153</td>
<td>5,423</td>
<td>12,800 (100%)</td>
</tr>
</tbody>
</table>

The evaluation metrics were the standard precision, recall and F1-score. A detected relationship was counted as true positive if the silver label set and model agreed on the existence of the relationship.
Table 3: Definition of comparative baselines; specific settings and relationship collection statistics included.

<table>
<thead>
<tr>
<th>Setting/Statistic</th>
<th>CO-OAS-BT</th>
<th>CO-OAS-BF</th>
<th>Baselines</th>
<th>SR-OAS</th>
<th>SR-MED</th>
<th>PAS-MED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus Filter</td>
<td>OA-Subset2014</td>
<td>OA-Subset2014</td>
<td>OA-Subset2014</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Concept Extraction Method</td>
<td>QuickUMLS*</td>
<td>QuickUMLS</td>
<td>SemRep**</td>
<td>SemRep</td>
<td>MetaMap***</td>
<td></td>
</tr>
<tr>
<td>Concept Filter:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Best Match Setting</td>
<td>true</td>
<td>false</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>- Similarity Filter</td>
<td>&gt;0.7</td>
<td>&gt;0.7</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Relation extraction method</td>
<td>co-occur</td>
<td>co-occur</td>
<td>SemRep</td>
<td>SemRep</td>
<td>PasMed</td>
<td></td>
</tr>
</tbody>
</table>

Test Set Statistics:
- # PMID used: 675,052 (CO-OAS-BT), 20% was set aside for testing. The best average F1 score across the five folds for a single epoch was then used for comparison.
- # distinct disorders: 29,298, 52,434, 10,826, 158,674, 108,822.
- # distinct Symp-Dis rels: 353,638, 1,733,461, 2,576,597.
- # distinct Treat-Dis rels: 333,493, 1,848,135, 11,436.
- # distinct Test-Dis rels: 133,199, 776,312, 11,436.

*QuickUMLS* version 1.2 built on UMLS dataset 2018AA; **SemMed v3.1** using SemRep v1.7. Predication table containing the relationships processed Medline articles up to December 31, 2017; ***MetaMap** v2016v2 with UMLS dataset 2018AA.

For the LSTM model comparison with the existing models, five fold cross validation was employed where 80% of the label data was used for training and 20% was set aside for testing. The best average F1 score across the five folds for a single epoch was then used for comparison. Cross-fold validation was not employed for trialling and comparing the LSTM models and features, however training and test data was divided using the same 80/20 split.

**Instantiation of Existing Models For Detecting Disorder-Centric Biomedical Relationships**

The proposed LSTM model is compared with a number of existing models; these baselines were selected according to the following criteria: (a) Relationships are extracted from the Medline corpus and/or the TREC 2014 document collection, consisting of the Open Access Subset of Pubmed Central, taken on January 21, 2014, herein called OA-Subset2014. This subset of Medline was selected because it is well defined, commonly used and easily accessible to researchers; (b) The extracted relationship types must include at least the four disorder-centric relationships targeted in this study; (c) The extracted biomedical relationships must be expressed as a relationship between two UMLS concepts so that evaluation between extraction methods can take place in the UMLS concept space. Table 3 defines each baseline model, the settings employed for each and the statistics of the resulting test collection. Figure 3 defines the generic pipeline employed for preparing each model to generate the test collections.

**Figure 3:** Pipeline Diagram with the text along the top describing the steps used to extract relationships and the text along the bottom providing examples of each step

For all models, a semType filter was applied, matching the disorder-centric concept classifications specified in Table 1. In addition, because SemRep also included a fixed set of predicates, we limited the predicates to those appropriate for each relation (e.g., disorder-treatment (DT) relationships were limited to treats, neg_treats predicates). (For the sake of brevity, these are not listed here, however are available upon request.) Generic, high-use concepts, such as treatment, patient, disease were removed prior to labelling, and therefore not tested in our evaluations.
In addition to the settings identified in Table 3, the following pre-processing steps relate to specific models. For the co-occurrence models, a text window size of 5, within sentence was used to capture two concepts within a relationship. The BestMatch setting relates to the CUI(s) selected within the co-occur window. When bestMatch=false, all possible concepts extracted from the window of text are included within the window, i.e. often more than one concept for the same text and concept borders can overlap, whereas when BestMatch=true, concept borders can not overlap and only the best concepts are selected. PASMED relationships were extracted by firstly filtering for the correct semType combinations, then MetaMapping both of the medical terms in each relationship to identify the CUIs and filtering across only those CUIs within the label data set. Because an actual relationship type was not provided, but rather the verb relating the two CUIS, all relationships were considered valid if the CUIs formed a valid disorder-centric semType pair, irrespective of the order of the CUIs. This provided PASMED results with optimistic recall.

Results and Discussion

LSTM Model Selection Results

Table 4 lists the model and feature definition changes and their impact on the extraction system performance. The changes made at each step are cumulative, such that each model builds upon the model and feature set of the previous model. The cumulative impact on F1-score is graphed in Figure 4.

The baseline model consists of a single LSTM layer and output RNN with input features including the location, cosine similarities, POS of a single pre-word to the CUI and CUI embeddings\(^\text{23}\) (500 dimensions). Improving the input features had a smaller overall impact (16.1%) on F1-scores, when compared with improving the model (38.6%). Incorporating a dense layer, i.e., a fully connected feed forward layer, at the output of the second LSTM, was the most important model improvement, adding 21.6% to F1 with all of the benefit arising from a 58.5% improvement in recall. The intuition behind the addition of the dense layer was to introduce a learning layer at a higher level of abstraction that might identify patterns across the output of the LSTM units. In particular the authors speculate that because of the wide variation in input sequences, the dense layer can take this variability into account to identify enable the identification of more positive relationships. The most important feature addition was the CUI semType embedding, improving recall, over the previous model, by almost 15%, resulting in the best overall, recall model (R=70.3%). No available learned semType embeddings from Medline were available, and so one-hot vectors were employed. A learned embedding of semTypes is likely to result in further gains, which was left for future work.

<table>
<thead>
<tr>
<th>Id</th>
<th>Model Definition</th>
<th>Feature Definition</th>
<th>Precision (%)(^*)</th>
<th>Recall (%)(^*)</th>
<th>F1-score (%)(^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>M1=Baseline LSTM</td>
<td>F1=Baseline features</td>
<td>0.4467 (0.0)</td>
<td>0.3154 (0.0)</td>
<td>0.3697 (0.0)</td>
</tr>
<tr>
<td>S1</td>
<td>M1</td>
<td>F2=F1+10 pre-word POS</td>
<td>0.4347 (-2.7)</td>
<td>0.3442 (+9.1)</td>
<td>0.3842 (+3.9)</td>
</tr>
<tr>
<td>S2</td>
<td>M2=M1+extra LSTM layer</td>
<td>F2</td>
<td>0.4457 (+2.5)</td>
<td>0.3856 (+12.0)</td>
<td>0.4134 (+7.6)</td>
</tr>
<tr>
<td>S3</td>
<td>M3=M2+dense layer</td>
<td>F2</td>
<td>0.4271 (-4.2)</td>
<td>0.6113 (+58.5)</td>
<td>0.5028 (+21.6)</td>
</tr>
<tr>
<td>S4</td>
<td>M3</td>
<td>F3=F2+semType</td>
<td>0.4463 (+4.5)</td>
<td>\textbf{0.7024} (+14.9)</td>
<td>0.5458 (+8.5)</td>
</tr>
<tr>
<td>S5</td>
<td>M3</td>
<td>F4=F3+new CUI2VEC</td>
<td>0.4799 (+7.5)</td>
<td>0.6895 (-1.8)</td>
<td>0.5659 (+3.7)</td>
</tr>
<tr>
<td>S6</td>
<td>M4=M3+semMed training data</td>
<td>F4</td>
<td>\textbf{0.6185} (+28.9)</td>
<td>0.6395 (-7.2)</td>
<td>\textbf{0.6191} (+9.4)</td>
</tr>
</tbody>
</table>

Introducing positive semMed\(^\text{14}\) training data had the second highest impact (9.4%) on model performance and the greatest positive impact on precision across all changes. It is widely recognised that training deep learning models on imbalanced label sets can adversely impact model performance. The silver label data set contains 2.4 times more negative than positive labels (Table 2). To balance the training data set, disorder-centric semMed relationships were selected from source documents outside of the silver label set but within the OA-Subset2014 collection, until the positive/negative label ratio was equal. The balanced and larger training set improved precision by 28.9%, however it is unclear what proportion of this change is due to the larger training set or better label balance. Training the model with a balanced silver data set would reduce its size by 41%, thereby rendering the results equally unclear and semMed only provides positive labels, so balance cannot be achieved by adding equal positive/negative labels. Better understanding the causal elements for this improvement has been left for further investigation.
Although the input features had a lesser impact on performance, each feature added at least 3.7% to the F1 score. The most important feature, CUI embeddings, were incorporated into the baseline model. Switching to our custom embedding, improved the F1 score by 3.7%, however this may have been a result of the expanded training base. The custom embedding supported more than 5 times the number of CUIs enabling a 39% increase in the training set size. 

Like the addition of semMed training data, the performance increase was the net result of improved precision and decreased recall (+7.5% vs -1.8%). It is unknown whether the improvement was due to the increased training set or the change in CUI embedding used. Understanding the relative benefit would be helpful and was left for future work. Also left for future work was replacing the CUI embedding with a deep, pre-trained model, such as those used in language model or text classification (E.g., ELMo, ULMFiT) which incorporate contextual information beyond that of the semantic embeddings used in this work.

**Comparative Results**

Table 5 lists the results by test system evaluated on the silver label data set. The label collection was derived from 754 documents within the OA-Subset2014 citation collection so that all systems tested on this corpus have equal access to the source documents.

The LSTM, inter-sentence relationship detection model achieved the best performance, compared with all systems, whether they utilised the whole of Medline or the OA-Subset2014 collection to derive their relationship sets. It achieved improvements of +82% in precision, of +57% in recall and of +78% in F1-score, when compared with the best corresponding result for any other system. Three limitations to the evaluation warrant mention here. First, the label collection quality is classed as ‘silver’: the detection of relationships is done by non-experts assessing each sentence from an English grammar standpoint, rather than a medical one. The use of multiple labelers and multiple validation paths for each label is designed to minimise poor labels, however until the data set is validated by a domain expert, errors may exist. Explication and validation of this silver dataset is the subject of future work. Second, at this point it is not possible to confirm that the superior performance of the LSTM model is due to the ability to detect more relationships, via inter-sentence relationship capture, or other factors. Analysis of the label collection reveals that 58% of the labels are inter-sentence, which might favour this hypothesis, however confirming this is left to future research. Third, the LSTM model was trained via 5-fold cross validation on the label collection. While over-fitting may occur for such a method on a smaller dataset, the standard deviation of F1 scores across five folds was 0.0307, which was 5.0% of the F1 average, suggesting minimal over-fitting.

**Table 5:** Performance results for each system based on the data collection.

<table>
<thead>
<tr>
<th>Test System</th>
<th>Data Collection</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO-OAS-BF</td>
<td>OA-Subset2014</td>
<td>0.2966</td>
<td>0.4146</td>
<td>0.3458</td>
</tr>
<tr>
<td>CO-OAS-BT</td>
<td>OA-Subset2014</td>
<td>0.2933</td>
<td>0.3493</td>
<td>0.3189</td>
</tr>
<tr>
<td>SR-OAS</td>
<td>OA-Subset2014</td>
<td>0.3400</td>
<td>0.1037</td>
<td>0.1589</td>
</tr>
<tr>
<td>SR-MED</td>
<td>MEDLINE</td>
<td>0.3218</td>
<td>0.2735</td>
<td>0.2957</td>
</tr>
<tr>
<td>PAS-MED</td>
<td>MEDLINE</td>
<td>0.3103</td>
<td>0.3976</td>
<td>0.3486</td>
</tr>
<tr>
<td>LSTM-DL</td>
<td>NA</td>
<td>0.6162</td>
<td>0.6487</td>
<td>0.6189</td>
</tr>
</tbody>
</table>

Between the test systems that were constructed with the OA-Subset2014 citation collection, the co-occurrence model with bestMatch=false produced the best F1 score (0.3458) with the highest recall (0.4146). It was hypothesised that this model would generate higher recall than the alternative bestMatch=true model because of the increased number
of CUIs and therefore relationships that are identified. However, it was not expected to produce similar, or even higher precision. This result is possibly explained by the quality of CUI mapping, which despite the setting name, bestMatch, struggles to select the correct CUI for a set of medical text. By using the BestMatch=false, more CUI options are created within the selection window, generating more correct relationships and because of the disorder-centric filter, less of the invalid relationships are kept. Although semRep generated particularly low recall results (0.1037), the relationships that were identified were more likely to be correct than in the co-occurrence models, reflected by a +14.6% improvement in precision over the best co-occurrence model precision.

Expanding the source of relationships from the OA-Subset2014 citation collection to the whole of Medline improved the SemRep system F1 performance by +86%, solely through a +164% increase in recall. PASMED, which demonstrated a +71% increase in recall over SemMed in Nguyen et al.’s work, showed a +45.5% increase in this comparison. The lower increase could be a result of the specific disorder-centric nature of the relationships extracted in this work. Overall, the PASMED system produced the best F1 score of all the existing baselines.

All of the existing baselines utilise within-sentence relationship extraction methods and the maximum recall achieved by any system was 0.4146, which would indicate that many of the labelled relationships were indeed inter-sentence. This helps to explain why the LSTM model could achieve a step-wise improvement in relationship-detection, over these existing baselines. This hypothesis could be tested in future work by extending the co-occurrence window to citation level co-occurrence, which should capture all relationships. If many relationships are cross-sentence, as indicated, then the disorder-centric LSTM model may have broader application in other BRE domains where relationships also occur between sentences.

Conclusion

In this paper we introduce a deep learning model to detect disorder-centric relationships between medical concepts across sentence boundaries. Relationships between disorders and their associated tests, treatments and symptoms underpin essential information needs of clinicians and can support biomedical knowledge bases, information retrieval and ultimately clinical decision support. The task was re-cast as a sequential labelling problem that could be tackled by a deep learning LSTM model approach. An inter-sentence, disorder-centric silver quality label collection was created by non-expert humans for training and evaluation purposes.

The cumulative contributions of word POS, punctuation and semType embeddings are compared together with important model changes. The resulting solution is then tested against other existing broad, biomedical relationship extraction methods, including co-occurrence, semRep and PASMED. The LSTM model outperforms these other models by at least +78% in F1 score suggesting that (a) inter-sentence relationships form an important component of disorder-centric relationships in biomedical literature and (b) the proposed LSTM deep-learning model and input sequence definition is a suitable approach for extracting these relationships. Although the LSTM model tested here was a suitable initial investigation, the next step is to consider other more recent deep learning models, for example transformer and end-to-end learning models, to compare the impact on BRE.

We also explore the impact of concept extraction methods (bestMatch = true/false) and collection size (Medline versus a 700K subset of Medline) on relationship extraction performance. Corpus size improves recall with a reduction in precision and using a broad concept selection policy (bestMatch=false) improves recall without deteriorating precision.

In conclusion, we provide a promising, new approach to inter-sentence, disorder-centric biomedical concept relationship extraction with noted limitations and numerous avenues to expand the technique to other biomedical domains.

References


Implementing the VMC Specification to Reduce Ambiguity in Genomic Variant Representation

Michael Watkins, Shawn Rynearson, Alex Henrie, and Karen Eilbeck
Biomedical Informatics, 421 Wakara Way, University of Utah, Salt Lake City, Utah 84108

Abstract
Current methods used for representing biological sequence variants allow flexibility, which has created redundancy within variant archives and discordance among variant representation tools. While research methodologies have been able to adapt to this ambiguity, strict clinical standards make it difficult to use this data in what would otherwise be useful clinical interventions. We implemented a specification developed by the GA4GH Variant Modeling Collaboration (VMC), which details a new approach to unambiguous representation of variants at the allelic level, as a haplotype, or as a genotype. Our implementation, called the VMC Test Suite (http://vcfclin.org), offers web tools to generate and insert VMC identifiers into a VCF file and to generate a VMC bundle JSON representation of a VCF file or HGVS expression. A command line tool with similar functionality is also introduced. These tools facilitate use of this standard—an important step toward reliable querying of variants and their associated annotations.

Introduction
As we near twenty years since the completion of the human genome and ten years since the breaking of the sequencing cost barrier, the amount of genomic variant data which has been generated and made available for research and clinical use is staggering. These variants are represented using nomenclatures and file types which are designed with enough extensibility to account for the growing complexity of variant data.

The most common variant file type is the Variant Call Format (VCF), which stores all variants from a single reference sequence in a tab-delimited format based on genomic coordinates. The most common variant representation nomenclature is called HGVS after the Human Genome Variation Society that develops it. This nomenclature specifies how to represent individual instances of genomic variation in a way that is human-readable.

Although typical analyses using variant data utilize both HGVS and VCF, there are a few fundamental differences between the two methodologies. For example, each methodology uses different variant normalization methods. Normalization refers to the position of a variant in the context of a nucleotide repeat—they can be left or right shifted. HGVS requires variants to be right-justified while VCF requires left-justification. This is further complicated as both HGVS and VCF continue to evolve to accommodate the growing complexity of variants. Versioning differences between tools implementing HGVS or VCF cause ambiguity in how these variants are represented, leading to complicated querying and annotation in later use.

Research institutions have been able to adapt their individual methods in response to this ambiguity. However, this has not been the case for clinical institutions that rely on strict informatics standards. Reproducible variant research and full clinical utilization of variant data all require a fundamental change in variant representation methods to address ambiguity. Individuals and institutions must be able to exchange variants and communicate about them with a surety that they are referring to the same genomic change on the same genomic reference sequence.

Ambiguity in Variant Representation
Ambiguous variant representation, caused by inconsistency in syntax structure, results in two or more representations of a single variant. Even the change of one letter in the syntax used to represent a variant can cause inconsistency and result in ambiguous representations. This causes downstream complications in associating annotations with that variant. Here we outline some causes of ambiguity in two popular variant representation methods.
VCF
The VCF standard has de facto become the principal variant file format in clinical practice, surpassing others such as genomeVCF and the genome variation format (GVF)\(^3\). This is more likely due to its long history and vast set of supporting software tools than to its ability to avoid ambiguity. In fact, the VCF specification is extremely flexible with many optional parameters (local phasing information, reference calls, no-calls, quality, likelihood, etc.) and many recommended, but not required, parameters (accession version, VCF version and HGNC gene identification)\(^3,\)\(^13\).
While this flexibility allows VCF to be used for a wide variety of applications, it creates two problems when translating to clinical space. The first is that clinical decision support applications require reliable data fields in the source data which they compute over and the variable nature of most of the fields in a typical VCF file leaves it unsuitable\(^12\). The second is ambiguity in the representation of the individual variants themselves. A simple example is acknowledged in the latest VCF specification\(^3\) and is shown in Figure 1.

Figure 1. Different representations of a two base genomic deletion.

Though difficult to detect at first, this entry shows an allelic variant which could occur at two different locations on the reference sequence. While this example is a simple two base pair difference in location, it would result in redundant entries to a variant archive and would complicate future extraction and annotation of the variant.

There are many more instances of ambiguity in the representation of variants which are possible because of the flexibility of the VCF standard. Most arise unexpectedly and, as mentioned previously, are accounted for in the research space by adjusting methods and processes in a way which is not feasible in clinical application.

HGVS
The HGVS nomenclature has been designed to describe the wide array of sequence variation including not only sequence changes, but biological mechanism, predicted events, and complex states\(^14\). There is an important distinction to be made between variant representation and variant annotation. A variant representation captures information related to the position of the variant with regards to a particular reference, where as an annotation refers to additional information such as interpretations made about the variant. Annotating a particular variant depends very much on the unambiguous representation of that variant. One study demonstrates the difficulty for several institutions to fully implement the HGVS nomenclature correctly in representing variants\(^15\), with many other studies showing downstream complications in annotating variants represented by ambiguous HGVS representations in variant archives\(^16\)-\(^22\).

A study by Yen et al. (2017) identified ambiguous HGVS representations both generated by certain tools and found in different variant archives. The study compared them to the preferred HGVS representation which was compiled using the most recent HGVS specification version. These examples are included in Figure 2 and show the ambiguity and discord which results from tools and archives not being able to keep up with a constantly evolving nomenclature. Interestingly, the study also identified ambiguous representations which each comply with the most recent HGVS specification version. These are shown in the last two rows of Figure 2 and demonstrate that ambiguity is an issue in the nomenclature itself and not just a result of differing versions being used by tools and archives. This issue must be addressed before reliable and standardized computational approaches to using variant data can be implemented.

<table>
<thead>
<tr>
<th>CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>4</td>
<td>.</td>
<td>GCG</td>
<td>G,GCGCG</td>
<td>.</td>
<td>PASS</td>
<td>DP=200</td>
</tr>
</tbody>
</table>

Possible ref sequences

\[ a \ t \ c \ G \ C \ G \ - \ - \ a \]
\[ a \ t \ c \ G \ - \ - \ C \ G \ a \]
Ambiguous HGVS representations | Source
--- | ---
c.2331_2339dup | Preferred HGVS representation
 c.2339_2340insGGGCTCCCC | COSMIC (archive)
c.2331_2339dupGGGCTCCCC | SnpEff (tool)
c.1895 + 5_1895 + 8delGTGA | Preferred HGVS representation
 c.1895 + 1_1895 + 4delGTGA | ClinVar (archive)
c.1895 + 9GTGAC > C | VR (tool)
p.Arg227Lysfs | Preferred HGVS representation
 p.Arg227fs | Preferred HGVS representation
 p.Arg227LysfsTer31 | Preferred HGVS representation
p.Gln100* | Preferred HGVS representation
p.Gln100Ter | Preferred HGVS representation

Figure 2. Examples of multiplicity of variant representation within HGVS nomenclature.

**General sources of ambiguity**
While fundamental aspects of VCF and HGVS make them particularly prone to ambiguous variant representation, there are additional more generalized sources of ambiguity that vary between tools and archives. These include whether coordinates are 0-based (as they are in the UCSC genome browser\textsuperscript{23}) or 1-based (as they are in the Ensembl genome browsers\textsuperscript{24}), which reference sequence is used, and whether or not the coordinates used are inclusive or interbase.

**Clinical effects of ambiguous variant representation**
Two independent studies have produced technical desideratas for the integration of genomic data into a clinical setting\textsuperscript{8, 10}. These describe a future of clinical systems which leverage information from many different genomic and non-genomic data sources in order to increase the ability of clinicians to use genomic data to make relevant treatment changes or recommendations. However, the studies conclude that this ideal is only possible through certain technical changes in how genomic data is represented and stored. For example, the actual variant data should be separated from clinical observations and should support lossless data compression\textsuperscript{8}. Variants should also be able to be classified into groups of common clinical impact while still supporting the ability to reference at the individual variant level when necessary\textsuperscript{10}. One desiderata also calls for a common knowledge base which is deployed at and developed by multiple independent organizations. This would allow affordable access to comprehensive genomic data\textsuperscript{10}. However, such a federated approach would likely have a high tolerance for redundant submissions and ambiguous entries unless a more reliable representation standard is adopted.

Although the opportunity for clinical genomic innovations and tools to provide recommendations and decision support to clinicians is constantly growing\textsuperscript{9, 25-27}, their implementation is stymied by the issue of ambiguity in variant representation\textsuperscript{8, 29}. Mismatched variant annotations and subsequent false treatment changes could be disastrous\textsuperscript{10}. These issues have begun to be addressed by a GA4GH task force, which assigned a sub-group, the VMC, to develop a variant model and a defined specification to facilitate variant representation and reliable exchange.

**Variant Modelling Collaboration**
The purpose of the Variant Modelling Collaboration (VMC) is to address the issue of unreliable genomic data exchange by developing a fundamentally different approach to representation. The workgroup developed new data models to represent not only allelic variations, but variation at the genotype and haplotype level. These models are straightforward and well-adapted to future modifications\textsuperscript{14}. The data models themselves are computationally digested
to generate unique identifiers to serve as a machine-readable representation of the variant. Thus, the VMC data model unambiguously names variation, with respect to reference sequence, and is well-equipped to foster reliable exchange between institutions or queries from external variant knowledge bases.

The VMC specification breaks variant data into fundamental components. These components are represented in specific data models and then run through the VMC digest algorithm (a combination of hash and encoding algorithms) to create unique identifiers which represent those specific components\textsuperscript{14}. The following is an example, shown visually in Figure 3, of creating a VMC identifier for an allelic variant:

1) Take the raw sequence data for the reference sequence and run it through the digest algorithm. This returns a VMC sequence identifier.
2) Take the start and end coordinates for the allelic variant and use them, according to the VMC interval model, to create a VMC interval object.
3) Run both the VMC sequence identifier and VMC interval object through the digest as a VMC location object, according to the model, to generate a VMC location identifier.
4) Use the VMC location identifier and the alternate base(s) found in the previously specified interval (referred to as state in the VMC specification) to create a VMC allele object according to the model. Then digest that object to generate the ultimate goal of a unique VMC allele identifier.

The current VMC specification also has data models to represent and digest identifiers for haplotypes and genotypes. A VMC haplotype is used to designate multiple VMC alleles as being in phase or “cis”. A VMC genotype is used to designate a group of VMC haplotypes. Both models are included in Figure 3.

<table>
<thead>
<tr>
<th>Object Type</th>
<th>Information Model</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>string, constrained by characters representing IUPAC nucleic acid or amino acid codes.</td>
<td>NNNN...ATGC...NNNN [63887120 bases]</td>
</tr>
<tr>
<td>Haplotype</td>
<td>&lt;Haplotype</td>
<td>completeness</td>
</tr>
<tr>
<td>Genotype</td>
<td>&lt;Genotype</td>
<td>completeness</td>
</tr>
</tbody>
</table>

**Figure 3.** Identifier generation for data types needed for variant representation.
standardized VMC data model digested into a unique identifier removes variant representation ambiguity and fosters reliable variant matching.

**Implementation**
We have developed a VMC Test Suite as an implementation of the schemas and digest algorithm explained in this first version of the specification. The Test Suite has four tools which have been made available as publicly accessible web tools hosted at http://vcfclin.org. The code can be found at https://github.com/eilbecklab/VMC-Software-Suite.

1) The first tool, shown in Figure 4, accepts a user-uploaded VCF file. It then goes through each variant of the file and pulls location intervals and alternate bases in order to generate a VMC allele identifier for each entry. It stores that identifier, along with the VMC sequence identifier and VMC location identifier used in the process, in the info field of that variant entry. It also adds requisite header lines which explain the additions. The modified file can then be downloaded back by the user.

2) The second tool also takes a user-uploaded VCF file as input. However, rather than give back a modified VCF file for download with VMC identifiers added in, the user is given a VMC bundle object, represented in the JSON format, which contains each of the VMC objects and identifiers generated from the file. The VMC bundle is envisaged as a mechanism to integrate complex variant data such as haplotypes into a workflow.

3) The third tool implements existing code found in the VMC GitHub repository which converts an HGVS string into a VMC bundle. This VMC bundle is a JSON object which holds each of the VMC identifiers generated for the variant encoded by the HGVS string.

4) A command line tool has also been developed with very similar functionality to each of the three tools. It is a lightweight implementation of the core data models and digest algorithm included in the VMC specification. It is publicly available through a GitHub repository.

---

**Figure 4.** One of the VMC Test Suite tools which adds VMC identifiers to a VCF file.
Evaluation

In order to confirm correct implementation of the VMC data models, digest algorithm, and output format, three validation tests were performed on the VMC Test Suite. These tests focused on the three main processes which enable the functionality of the tools in the suite.

1) The first test was to validate our process of generating sequence identifiers. SeqRepo is a large collection of biological sequences which is made available by Biocommons. The VMC developers were granted access to SeqRepo and generated a sequence identifier for each entry. We used SeqRepo as the gold standard for this test.

2) The second test was to validate our process of generating allele identifiers. Since allele identifiers require a location identifier (as shown in Figure 3), this would also test our process of generating location identifiers. The VMC developers created Python code which generates VMC identifiers for HGVS expressions. We used outputs of that code as the gold standard for this test.

3) The third test was to validate our implementation of the VMC bundle JSON schema. This schema is available on the VMC GitHub repository and was used as the gold standard for this test.

Test 1: Sequence identifiers

The tools in the suite each draw from a custom database of pre-digested sequence identifiers which grows as new variants are encountered. This cuts down on future processing time. The suite must use information from the VCF entry to download the appropriate reference FASTA file for that entry from NCBI. This FASTA file is then run through the VMC digest algorithm in order to generate the sequence identifier. Since a difference of even one base in the FASTA file used would result in non-identical sequence identifiers, our process of locating and digesting the appropriate FASTA file had to be validated. To do this, we simply compared the sequence identifiers in our database to those found in a gold standard database. Both associated an accession number with the sequence identifier, making the look-up process simple.

We found that, of the 1720 accession numbers which were shared between the databases, each one had identical sequence identifiers. There were 204,819 accession numbers in our database which weren't found in SeqRepo but since the process used to generate their corresponding sequence identifiers was the same as the 1720 matches, these can be assumed to be accurate sequence identifiers as well.

Test 2: Allele identifiers

As explained previously, VMC allele identifiers are each generated by digesting a data model which includes a VMC sequence identifier, a VMC location identifier, and an allelic change. After getting the right sequence identifier for a given VCF entry, the suite extracts all the information needed to fill the location and allele data models. It digests the location data model and uses it to complete the allele data model. This model is then digested into an allele identifier.

To test these processes, we needed a VCF file which had HGVS expressions for each entry. This would provide an HGVS expression to use as input for the HGVS conversion code written by the VMC developers. Allele identifiers generated with this code would serve as a gold standard. We would then be able to compare that identifier to the one generated by the suite (generated from the other fields of that VCF entry). Identical allele identifiers would show that the suite is generating allele identifiers (and, by extension, location identifiers) properly.

ClinVar provides a weekly release VCF file of their variants which have HGVS expressions included in the INFO field for each entry. An example entry from the file, with portions of the INFO field omitted to highlight the included HGVS expression and rsID, is shown in Figure 5. The included rsID was also important because it allowed the suite to select the appropriate sequence identifier.
### Figure 5. Entry from the ClinVar weekly release VCF file.

Of the 393403 variants in the file, the suite generated allele identifiers for 381257 of them (96.91%) which were identical to the gold standard allele identifiers. Of the non-identical allele identifiers, 11450 of them (2.91% of total) came from deletions, 682 of them (0.17% of total) came from insertions, and 14 of them (0.0036% of total) came from indels. On closer inspection, these non-identical allele identifiers represent an assortment of edge cases where the coordinates of the VCF entry doesn’t match the location in the corresponding HGVS expression.

The identical allele identifiers include a high percentage of all forms of variation accepted by the gold standard HGVS conversion code (substitutions, insertions, deletions, and indels). We believe this test validates our approach for generating correct VMC identifiers and points to rare situations where VCF and HGVS are not aligned.

**Test 3: VMC bundle JSON**

To ascertain the JSON representation of the results of the tools was true to the VMC bundle JSON schema included in the VMC specification, we used the jsonschema Python tool to check each field. This tool confirmed that the output of our second VCF tool (creates a VMC bundle in JSON from a VCF file) matches the schema.

**Discussion**

The VMC specification provides an important building block in the effort to overcome ambiguity in variant representation and resulting discordance in variant archives. It also provides a backbone for naming complex variants such as haplotypes, a feature necessary for communication of pharmacogenomic and immunologic variants. It takes a fundamentally different approach to the process of representing variants and provides unique identifiers to enable reliable querying and annotation of those variants. However, the next important problem which must be considered is that of equivalence.

General sources of ambiguity such as different reference sequences, different coordinate systems, alternate transcripts, etc., do not go away with the VMC data models. However, VMC does provide more stringent representation standards and as a result, the concept of equivalence definitions is now a possibility. Four specific types of equivalence are discussed here with the acknowledgement that additional considerations will continue to become more apparent as the VMC standard is adopted.

**Normalization**

Different normalization strategies and tools can result in slightly different location intervals. These differing location intervals could lead to different VMC identifiers being generated which should be considered equivalent. Anticipating this, the VMC specification requires that the VT normalization algorithm be used to normalize variants prior to their inclusion in the data model and digestion into a VMC identifier. By requiring all variants to be normalized using the same algorithm, it will be possible to computationally determine normalization equivalence and account for it in future implementations of the VMC specification.

**Projection**

A variant may be represented on a genomic sequence, on a transcript, or on a protein sequence. With the volume of genomic data now available it is likely that any particular variant could be represented in all three ways. Location
coordinates and reference sequences will be different for each representation but, in the end, they are referring to the same variant and should be considered equivalent. This would allow the combination and utility of all three levels of annotation associated with those sequence types.

Alternate transcripts
A common source of ambiguity is different variant locations arising from alternate transcripts. Biologically, the different splice sites of a particular genomic region can result in a variant being found on two non-identical transcripts. While the transcripts themselves are non-identical, the variant is functionally the same. The different transcripts lead to different variant locations which ultimately would result in different VMC identifiers. However, these different identifiers should be considered equivalent because they refer to a functionally-equivalent variant. Establishing this equivalency would require computation over gene annotations and genomic-level interpretations. While the computation would be complex, VMC makes this equivalence possible to determine in future implementations.

Lift-over
The idea of lift-over is purely systematic and not biological. It refers to the ability to take a variant backward and forward between different genome builds and versions while maintaining equivalence. Each genome build and version will generate a different VMC identifiers which are functionally identical and should be considered equivalent. NCBI currently hosts a lift-over tool called Remap which projects annotation data between different coordinate systems\(^{35}\). Providing this functionality for variant representations on a larger scale will be a possibility as VMC implementation continues to grow and many more VMC identifiers are generated and associated.

<table>
<thead>
<tr>
<th>Normalization</th>
<th>Alternate transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Normalization Diagram" /></td>
<td><img src="image2" alt="Alternate Transcripts Diagram" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Projection</th>
<th>Lift-over</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="Projection Diagram" /></td>
<td><img src="image4" alt="Lift-over Diagram" /></td>
</tr>
</tbody>
</table>

**Figure 6.** Example depictions of the four types of equivalence discussed here.

**Conclusion**
We have presented an implementation of the GA4GH VMC specification, with the hope that these tools will facilitate the use of this variant representation standard. Using VMC identifiers to represent variant data will reduce redundancy within variant archives. The computationally stringent implementation requirements will also significantly reduce discord among tools used to generate VMC identifiers. This reduction in ambiguity in variant representation will allow more reliable and precise data queries, resulting in more reproducible research methods and many useful clinical applications which were not possible before.
Acknowledgments
GA4GH is increasingly involved in standards development for biological data. The VMC group provides an open environment for discussion and development of specifications and tools for variant representation. We are thankful for the productive weekly discussion and innovation. We want to highlight Reece Hart for initiating this group and for his leadership. This work was supported by the National Institute of Health: R01HG008628 to KE, and NLM T15-LM007124 training predoctoral slot to MW.

References
23. UCSC Genome Browser Home. Available at: https://genome.ucsc.edu/.
30. VMCC! GitHub Repository. Available at: https://github.com/srynobio/vmcl.
31. SeqRepo GitHub Repository. Available at: https://github.com/biocommons/biocommons.seqrepo.
Relation Extraction from Clinical Narratives Using Pre-trained Language Models

Qiang Wei, M.S.¹, Zongcheng Ji, Ph.D.¹, Yuqi Si, M.S.¹, Jingcheng Du, B.S.¹, Jingqi Wang, M.S.¹, Firat Tiryaki, B.S.¹, Stephen Wu, Ph.D.¹, Cui Tao, Ph.D.¹, Kirk Roberts, Ph.D.¹, Hua Xu, Ph.D.¹
¹School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract
Natural language processing (NLP) is useful for extracting information from clinical narratives, and both traditional machine learning methods and more-recent deep learning methods have been successful in various clinical NLP tasks. These methods often depend on traditional word embeddings that are outputs of language models (LMs). Recently, methods that are directly based on pre-trained language models themselves, followed by fine-tuning on the LMs (e.g., the Bidirectional Encoder Representations from Transformers (BERT)), have achieved state-of-the-art performance on many NLP tasks. Despite their success in the open domain and biomedical literature, these pre-trained LMs have not yet been applied to the clinical relation extraction (RE) task. In this study, we developed two different implementations of the BERT model for clinical RE tasks. Our results show that our tuned LMs outperformed previous state-of-the-art RE systems in two shared tasks, which demonstrates the potential of LM-based methods on the RE task.

Introduction
Electronic Health Records (EHRs) contain massive free-text data documented by healthcare professionals during the course of patient care, such as clinical notes, discharge summaries, lab reports, and pathology reports¹. Compared to structured data, free-text data often records detailed information of clinical events and communications between teams in hospital settings. Much important patient information often exists in unstructured format only, and manual extraction of such information is expensive and time-consuming. Therefore, applying natural language processing (NLP) technologies to extract related patient information from clinical notes is highly recommended to support clinical research and applications².

Relation extraction (RE), one of the essential tasks of information extraction, aims to identify semantic connections between mentions of concepts in a document³,⁴. For example, in the sentence “an MRI revealed a C5-6 disc herniation with cord compression,” we attempt to identify that the test “MRI” reveals two medical problems “a C5-6 disc herniation” and “cord compression”. Previous studies have investigated diverse types of relations, such as disease-attribute pair extraction⁵,⁶, temporal relation identification⁷, adverse drug event detection⁸,⁹, etc. Recently, the clinical NLP community started a series of shared tasks on relation extraction from clinical notes, including the Informatics for Integrating Biology and the Bedside (i2b2) challenges⁷,¹⁰, the Semantic Evaluation (SemEval) challenges¹¹,¹², the BioCreative V task¹³, and the most recent 2018 National NLP Clinical Challenge (n2c2)¹⁴. These open challenges greatly facilitate the development resources (e.g., corpora) and methods for RE in the medical domain.

Early RE systems can be classified into two major categories: (1) rule-based methods; and (2) machine learning-based methods. Rule-based methods such as dependency trees¹⁵ and coreference chains¹⁶ are used to extract relations. Machine learning-based methods such as Support Vector Machine (SVM)¹⁷ and Conditional Random Field (CRF)⁴ are widely applied. However, current existing RE methods largely rely on heavy preprocessing steps that require extracting features from text, such as lexical, syntactic and semantics information¹⁸. Those extraction at early steps might introduce sources of errors that would accumulate to the following RE task¹⁹. These limitations severely restrict the portability and generalizability of RE to other novel resources.

In the past few years, deep learning methods have demonstrated their effectiveness and often achieved the state-of-the-art in diverse NLP tasks. Sequential-based deep learning methods including recurrent neural networks (RNN), convolutional neural networks (CNN) and their variants have been applied to clinical relation extraction²⁰,²¹. More recently, Bidirectional Encoder Representations from Transformers (BERT)²² introduced masked language models and reported superior performance on multiple benchmark datasets of diverse NLP tasks. BERT is pre-trained on a large corpus in an unsupervised manner and can then be fine-tuned on a downstream task with a simple layer on top of its architecture. By fine-tuning as a whole, the network adjusts the entire language model and thus encodes more contextualized information. BERT has shown its benefits in downstream NLP tasks such as concept extraction²³,²⁴, text classification²⁵, question answering²⁶, and text generation²⁷, especially pre-trained on a large clinical or
biomedical corpus. Due to their ability to capture contextualized information, deep language representation approaches achieve a task-specific architecture by simply appending the deep learning model for the downstream task.

Despite BERT’s success in the open domain and biomedical literature, it has not yet been applied to clinical RE tasks. In this study, we aim to investigate how to apply BERT to clinical RE tasks by (1) comparing two different implementations of the BERT model for clinical RE; (2) evaluating BERT models across different clinical RE tasks; and (3) assessing four BERT models trained from the open domain, biomedical literature, and clinical text on the clinical RE task. To the best of our knowledge, this is the first study to apply pre-trained language models of BERT to clinical RE tasks and our evaluation results show that it outperformed the previous state-of-the-art systems in clinical RE.

Material and Method

Tasks and Datasets

In this study, we used two datasets: the 2018 National NLP Clinical Challenges (n2c2) corpus and the 2010 Informatics for Integrating Biology & the Bedside (i2b2) challenge corpus. All named entities (clinical concepts) were given, and the task here was to identify relations between entities.

The n2c2 corpus included 505 discharge summaries, which came from the MIMIC-III (Medical Information Mart for Intensive Care III) clinical care database. The corpus contained nine types of clinical concepts including drug and eight attributes (reason, frequency, ADE, strength, duration, route, form and dosage). The relations between drug and the eight attributes were also provided (statistics of relations are in table 1). The task is to recognize all relations between drugs and attributes. The training set included 303 discharge summaries and the test set included 202 discharge summaries.

The other dataset used in this study came from the 2010 i2b2 challenge relation extraction task, including 426 discharge summaries collected from 2 hospitals. The dataset is a subset of the original dataset used in the challenge, since the University of Pittsburgh Medical Center’s data is not available to the public and was removed from the original dataset after the challenge. Only 170 of the original 394 training documents and 256 of 477 test documents were available for download, so we combined and re-split them in this study. The dataset includes 3 types of entities (medical problem, lab test and treatment) and 8 types of relations between them (Table 1).

BERT-based relation extraction

Given entities annotated in sentences, the relation extraction task can be transformed into a classification problem. A classifier can be built to determine categories of all possible candidate relation pairs $(e_1, e_2)$, where entities $e_1$ and $e_2$ are from the same sentence. For the n2c2 dataset, we generated candidate pairs by pairing each of the drugs with each of the attributes in a sentence. For example, given sentence “Furosemide 10 mg IV ONCE Duration : 1 Doses”, there were five entities “Furosemide”, “10 mg”, “IV”, “ONCE” and “1”. All possible candidate pairs were (10 mg, Furosemide), (IV, Furosemide), (ONCE, Furosemide) and (1, Furosemide). For the i2b2 dataset, we generated candidate pairs by pairing each of the problems with the other entities: including problems, treatments and tests. Before we build the classifiers for relation extraction, all documents went through a pre-processing procedure that includes basic steps such as sentence boundary detection and tokenization, which were done using the CLAMP (Clinical Language Annotation, Modeling, and Processing) Toolkit.

In this study, we developed two BERT-based methods: Fine-Tuned BERT and Feature Combined BERT to determine relation categories for these candidate pairs, described below.

**Fine-Tuned BERT (FT-BERT)**

*Input representation*

In order to represent a candidate relation pair in an input sentence, we used the semantic type of an entity to replace the entity itself. For example, as described above, there were four possible candidate pairs in the sentence “Furosemide 10 mg IV ONCE Duration: 1 Doses”, and it would be transformed into four samples (Figure 1). Each of them contained one candidate relation pair. In Sample 1, two entities “Furosemide” and “10 mg” were replaced by their semantic type
labels “@Drug$” and “@Strength$”. Note that even if an entity contains multiple words, it is still replaced by one label.

**Table 1.** Statistics of relations in the n2c2 and the i2b2 corpora.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Example</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>n2c2</td>
<td>Strength → Drug</td>
<td>Patient has been switched to <em>lisinopril</em> 10mg 1 tablet PO QD. 10mg → <em>Lisinopril</em></td>
<td>10946</td>
</tr>
<tr>
<td></td>
<td>Duration → Drug</td>
<td>Patient prescribed 1-2 325 mg / 10 mg <em>Norco</em> pills every 4-6 hours as needed for pain. 4-6 hours → <em>Norco</em></td>
<td>1069</td>
</tr>
<tr>
<td></td>
<td>Route → Drug</td>
<td>Patient has been switched to <em>lisinopril</em> 10mg 1 tablet PO QD. <em>PO</em> → <em>lisinopril</em></td>
<td>9084</td>
</tr>
<tr>
<td></td>
<td>Form → Drug</td>
<td>Patient has been switched to <em>lisinopril</em> 10mg 1 tablet PO QD. <em>tablet</em> → <em>lisinopril</em></td>
<td>11028</td>
</tr>
<tr>
<td></td>
<td>ADE → Drug*</td>
<td>Patient is experiencing <em>muscle pain</em>, secondary to <em>statin</em> therapy for coronary artery disease. <em>muscle pain</em> → <em>statin</em></td>
<td>1840</td>
</tr>
<tr>
<td></td>
<td>Dosage → Drug</td>
<td>Patient has been switched to <em>lisinopril</em> 10mg 1 tablet PO QD. 1 → <em>lisinopril</em></td>
<td>6920</td>
</tr>
<tr>
<td></td>
<td>Reason → Drug</td>
<td>Patient prescribed 1-2 325 mg / 10 mg <em>Norco</em> pills every 4-6 hours as needed for <em>pain</em>. <em>pain</em> → <em>Norco</em></td>
<td>8578</td>
</tr>
<tr>
<td></td>
<td>Frequency → Drug</td>
<td>Patient has been switched to <em>lisinopril</em> 10mg 1 tablet PO QD. <em>QD</em> → <em>lisinopril</em></td>
<td>10344</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TrIP</td>
<td>Treatment improves medical problem</td>
</tr>
<tr>
<td>TrWP</td>
<td>Treatment worsens medical problem</td>
</tr>
<tr>
<td>TrCP</td>
<td>Treatment causes medical problem</td>
</tr>
<tr>
<td>TrAP</td>
<td>Treatment is administered for medical problem</td>
</tr>
<tr>
<td>TrNAP</td>
<td>Treatment is not administered because of medical problem</td>
</tr>
<tr>
<td>TeRP</td>
<td>Test reveals medical problem</td>
</tr>
<tr>
<td>TeCP</td>
<td>Test conducted to investigate medical problem</td>
</tr>
<tr>
<td>PIP</td>
<td>Medical problem indicates medical problem</td>
</tr>
</tbody>
</table>

*ADE: adverse drug event.*

**Original sentence**  
[CLS] Furosemide 10 mg IV ONCE Duration : 1 Doses

**Transformed samples**

1. [CLS] @Drug$ @Strength$ IV ONCE Duration : 1 Doses
2. [CLS] @Drug$ 10 mg @Route$ ONCE Duration : 1Doses
3. [CLS] @Drug$ 10 mg IV @Frequency$ Duration : 1Doses
4. [CLS] @Drug$ 10 mg IV ONCE Duration : @Dosage$ Doses

**Figure 1.** An example of transformed samples from an original sentence used in FT-BERT.
**FT-BERT model**

Devlin et al.’s BERT model was used, and a linear classification layer was added on top to predict the label of a candidate pair in sentential context (Figure 2). In detail, a classification token [CLS] was added at the beginning of a sentence, whose output vector was used for classification. As typical with BERT, we used a [CLS] vector as input to a classification layer. Then a softmax layer was added to output labels for the sentence.

![Figure 2. The architecture of FT-BERT. CLS represents the [CLS] token; W_i represents words in sentence; E_1 and E_2 represent replaced semantic type labels of entities.](image)

**Feature Combined BERT (FC-BERT)**

**Input representation**

Instead of replacing entities with semantic type labels, we used additional BIO tags to represent entities in sentence, where “B” represents the beginning of an entity, “I” represents other words inside an entity, and “O” represents all other non-entity words. Compared with the input representation in FT-BERT, the BIO representation keeps entity word information so that the model has more information to classify. Figure 3 shows how to use BIO tags to represent sample (1) from Figure 1. Both sentence words and their tags are used together as input for FC-BERT.

![Figure 3. An example of the input representation in FC-BERT. The traditional BIO (Beginning, Inside, Outside) tags are augmented with the semantic types of the respective tokens.](image)

**FC-BERT Model**

We utilized the BERT model to generate vectors for all words in sequence; in parallel to the BERT model, the BIO tag sequence of the sentence was represented in an embedding layer. The vectors for words and the vectors for tags were concatenated (according to the original index position) and then sent to a classification layer, which was a BLSTM neural network with attention. Output labels per sentence were obtained via softmax.
Baseline methods

We also included a few baseline methods for comparison with the BERT models. For the n2c2 dataset, two strong baseline methods, the CNN-RNN method and the JOINT method were used, which were developed by our group and achieved the best performance on relation classification task in the n2c2 challenge\textsuperscript{32}. For the i2b2 dataset, the baseline methods used were the Seg-CNN method from Luo et al.,\textsuperscript{33} and the SDP method from Li et al.,\textsuperscript{34} which both outperformed the best systems that participated in the i2b2 challenge. We also compare with Li et al.’s recent Seg-GCRN method\textsuperscript{35}, which improves upon the Seg-CNN results incrementally. The Seg-CNN and Seg-GCRN evaluations used the original 2010 i2b2 dataset (including data from the University of Pittsburgh Medical Center) to develop and evaluate their methods, while the SDP method (and our experiments) excluded the Pittsburgh data. All results of baseline methods came from their original paper without re-running.

Experiments

For each BERT-based method, we evaluated four pre-trained language models, namely, (1) uncased BERT-large\textsuperscript{23}, (2) cased BERT-large\textsuperscript{23}, (3) the BioBERT model that was pre-trained using PubMed Central full text articles and PubMed abstracts\textsuperscript{24} and (4) the cased MIMIC BERT model that was pre-trained using the MIMIC III dataset\textsuperscript{25}. The difference between uncased BERT-large model and cased BERT-large model was that the former converted all words into lower case and the latter did not. For the n2c2 dataset, the original training set was randomly split into a new training set and a development set of 242 and 61 documents respectively (about 4:1), and the original test set (202 documents) was still used as test set for evaluation. For the i2b2 dataset, we mixed the original training set and test set together, then randomly split it into new training, development and test sets (with a ratio about 3:1:1). The development set was used for optimizing parameters, and the test set was used for evaluation. Note that this split of data implies that our i2b2 results are not directly comparable with the published literature using the original corpus. The evaluation metrics used in this study were as follows.

\[
\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad (1)
\]

\[
\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (2)
\]

\[
F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)
\]

Model parameters. The parameters of BERT were not frozen during training. The parameter maximum sequence length for both of methods were 128. The hidden layer size for Bi-LSTM in FC-BERT was 100 and embedding size of tag in FC-BERT was 100. For other parameters, we used the default parameters in the BERT.
Results

Table 2 shows results of the FT-BERT/FB-BERT models and baseline methods using four different pre-trained models on the n2c2 dataset. On the n2c2 data all four FT-BERT models that used different pre-trained BERT models outperformed the baseline method CNN-RNN. Three of them (except the FT-BERT that used cased BERT) outperformed the JOINT method on F1 score (up to 0.0023). All FC-BERT models performed worse than baseline method JOINT, but the uncased BERT model was better than the CNN-RNN method. Among FT-BERTs that used different pre-trained models, the FT-BERT that used the MIMIC BERT model was better than the one that used other pre-trained BERT models on F1 score and precision. The pre-trained BERT model using biomedical literature (BioBERT) showed no improvement on the clinical datasets compared with the models trained on data from the open domain. This may be because the BioBERT model trained based on the original BERT-base model includes fewer parameters than the original BERT-large models used in the study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>FT-BERT</th>
<th>FC-BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CNN-RNN</td>
<td>JOINT</td>
<td>Cased</td>
</tr>
<tr>
<td>n2c2</td>
<td>P</td>
<td>0.9673</td>
<td>0.9715</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.8878</td>
<td>0.9079</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0.9258</td>
<td>0.9386</td>
</tr>
</tbody>
</table>

All FT-BERT models outperformed the baseline method SDP on the i2b2 dataset under all three metrics of precision, recall and F1 score (Table 3). Only the FT-BERT models that used uncased BERT and MIMIC BERT outperformed the other two baseline methods Seq-CNN and Seq-GCRN. All FC-BERT models were worse than baseline methods. Among the FT-BERT models, the one that used MIMIC BERT was still the best on the i2b2 dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>FT-BERT</th>
<th>FC-BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDP</td>
<td>Seq-CNN*</td>
<td>Seq-GCRN*</td>
</tr>
<tr>
<td>i2b2</td>
<td>P</td>
<td>0.7569</td>
<td>0.748</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.7303</td>
<td>0.736</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0.7434</td>
<td>0.742</td>
</tr>
</tbody>
</table>

*These two methods were developed and evaluated on the original dataset of the 2010 i2b2 challenge.

Table 4 shows the F1 scores of our methods on each category of the n2c2 dataset. All four FT-BERT models performed better on all categories than CNN-RNN. The JOINT was slightly better than FT-BERT on four categories (Strength→Drug, Frequency→Drug, Form→Drug and Route→Drug). Compared with JOINT, three of four FT-BERT models improved the performance on the most difficult categories, reason and ADE; the FT-BERT models that used MIMIC BERT performed best (0.7674 vs. 0.7552, JOINT; 0.8124 vs. 0.7871, JOINT). The FC-BERT models performed better than the CNN-RNN but worse than the JOINT on these two categories.
In order to compare our methods with previous studies, we also merged all eight types of relations into three. Relations between tests and problems (TeCP and TeRP) were merged as Test-Problem, relations between treatments and problems (TrCP, TrAP, TrWP, TrNAP and TrIP) were merged as Treatment-Problem, and relations between problems and problems were kept unchanged. The performances of our methods were calculated on the merged categories. Table 6 shows the results. On overall performance, FT-BERT was 0.01 higher than the best baseline method Seg-
BERT method. First, the FC \( BERT \) made a substitution for all entities. In contrast, FT \( BERT \) made the patterns clear by replacing all entities with their semantic types. For example, in sentence “… given nebs, IV solumedrol and dose of levofloxacin IV and admitted for copd exacerbation.”, FT \( BERT \) recognizes all relations correctly and FC \( BERT \) fails to recognize all of them. The fact that FC \( BERT \) uses all information including words may also cause the failure of recognizing some parallel structure. For example, in sentence “… Vancomycin / Cefepime for Staph …”, FC \( BERT \) can only recognize relation between Cefepime and Steph, but fails to recognize relation between Vancomycin and Staph.”. Second, in the FC \( BERT \) method, a random initialized embedding layer was used to represent BIO tags, which may be too simple to capture the difference and connections between BIO tags. A pre-trained vector may improve the performance of FC\( BERT \).

The MIMIC FT \( BERT \) model performed the best on both of these two datasets, but on the i2b2 dataset it was only slightly better than the uncased \( BERT \) model. It may be because the n2c2 dataset was a subset of the MIMIC III corpus, so that MIMIC \( BERT \) was trained on in-domain data. Both cased \( BERT \) and uncased \( BERT \) were pre-trained on corpora from the open domain, but the performance of the latter was better. It may be because the word shape feature doesn’t help for RE. The \( BERT \) model that pre-trained on biomedical literature only slightly improved the overall performance on two datasets. One possible reason for this was that it is pre-trained on \( BERT \)-base instead of \( BERT \)-large, and the language used in biomedical literature is different from the language used in clinical narratives.

This is just the first attempt to apply \( BERT \) to clinical RE and there are different aspects that can be further improved. Our study followed the traditional framework of classifying candidate relation pairs, using the FT \( BERT \) and FC-
BERT models. Because the architecture of the JOINT method showed good performance on the n2c2 challenge, in the future, it’s possible to improve performance by combining BERT with JOINT. Moreover, we plan to further evaluate pre-trained language models from clinical corpora, e.g., supplement additional medical vocabulary and train uncased MIMIC III models.

**Conclusion**

In this study, we developed and evaluated BERT-based methods for clinical relation extraction. Our results show our RE methods based on pre-trained language models outperformed previous state-of-the-art RE systems in two shared tasks. In addition, our evaluation shows using clinical data to pre-train BERT models can benefit clinical RE.

**Acknowledgement**

This work is supported by NLM 5R01LM010681, NLM R00LM012104, NCI U24 CA194215, NIGMS 5U01TR002062 and the Cancer Prevention Research Institute of Texas (CPRIT) Training Grant #RP160015.

**Conflicts of Interest**

Dr. Xu and The University of Texas Health Science Center at Houston have research-related financial interests in Melax Technologies, Inc.

**References**


Towards Interpretable Skin Lesion Classification with Deep Learning Models

Alec Xiang¹, Fei Wang, PhD²

¹Horace Greeley High School, Chappaqua, New York; ²Weill Cornell Medical College, New York City, New York

Abstract

Skin disease is a prevalent condition all over the world. Computer vision-based technology for automatic skin lesion classification holds great promise as an effective screening tool for early diagnosis. In this paper, we propose an accurate and interpretable deep learning pipeline to achieve such a goal. Comparing with existing research, we would like to highlight the following aspects of our model. 1) Rather than a single model, our approach ensembles a set of deep learning architectures to achieve better classification accuracy; 2) Generative adversarial network (GAN) is involved in the model training to promote data scale and diversity; 3) Local interpretable model-agnostic explanation (LIME) strategy is applied to extract evidence from the skin images to support the classification results. Our experimental results on real-world skin image corpus demonstrate the effectiveness and robustness of our method. The explainability of our model further enhances its applicability in real clinical practice.

Introduction

Skin disease is one of the leading causes for global disease burden. It is reported that about 85 million Americans (27% of population; more than 1 in 4 individuals) were seen by dermatologists in 2013 and more than 9,500 people in the U.S. are diagnosed with skin cancer every day¹,². The diagnostic criteria of the skin diseases usually involve visual inspection of the skin lesions from their dermoscopic images as the first step. One such example is melanoma, which is one of the deadliest cancers. The diagnosis of melanoma includes two steps: visual inspection and biopsy. Because of the invasiveness of biopsy to patients, the accuracy of visual inspection is crucial. ABCDE (Asymmetric Shape, Border, Color, Diameter and Evolution) has been a popular rule for visual screening of melanoma based on the geometric characterization of the skin lesions³. However, without the equipment of these clinical knowledge, melanoma patients may not be aware of the severity of the disease they have and thus miss the best timing for treating their conditions.

In recent years, because of the rapid development of both computer hardware and software technologies, a large volume of skin images has been collected and sophisticated deep-learning based models have been trained to perform automatic analysis of these skin images. These models hold great promise as screening tools for skin diseases because of their superior capabilities of image analysis. For example, Esteva et al. proposed to adapt the Google Inception v3 model for melanoma detection from skin images⁴,⁵. The model, with parameters fine-tuned on 130K skin images, achieved classification performance comparable to human dermatologists. Kawahara et al. proposed a Fully Convolutional Neural Network (FCNN) model to extract multi-scale features skin images and perform lesion classification and achieved an accuracy of 81.8% on a 10-class skin disease classification problem, while the best reported classification accuracy on the same dataset is 67%⁶. Bi et al. proposed to leverage the deep Residual Network (ResNet) to perform melanoma detection and achieved state-of-the-art performance⁷,⁸.

Despite the good quantitative performance of these existing models, there are still some challenges to be addressed.

1) Model robustness. Existing research usually adapted well-established deep learning models (e.g., Inception or ResNet) trained on general computer vision tasks. It is difficult to guarantee any single model can work consistently well on different skin lesion images.

2) Sample limitation. In most of the existing studies, the amount of available skin images for training the complicated deep learning model is not enough⁹,¹⁰. Thus, these models typically will need to be pre-trained on other large-scale image data sets and the limited skin images will be used for fine-tuning to facilitate convergence.

3) Decision interpretation. Only quantitative classification results would not be enough for decision support in real clinical practice. As we introduced above, the dermatologists have specific diagnostic criteria to follow when they make the diagnosis. Therefore, the model we developed should also be able to generate the necessary evidence to support the classification results.
With the above considerations, we propose a novel deep learning pipeline for skin lesion classification in this paper.

![Pipeline Diagram](image)

**Figure 1.** Deep learning pipeline for skin lesion classification

It is worthwhile to highlight the following aspects of our pipeline, as shown in Figure 1:

1) Rather than using a single model, we ensembled a set of deep learning models (including VGG16, DenseNet, Xception, and Inception-ResNet v2) that have demonstrated to be effective on different computer vision tasks. In this way, we can sufficiently leverage the strength of different individual models and enhance the overall model robustness. We have also investigated different ensemble methods such as simple majority voting and model stacking.

2) In order to augment the training image set, we adopted the condition-based generative adversarial networks (GANs) on top of the traditional image augmentation approaches such as rotation and resizing.

3) Local interpretable model-agnostic explanation (LIME) strategy is implemented to interpret final model predictions, so that image features that lead to the final classification are highlighted.

We evaluate the effectiveness of the proposed pipeline using the Human Against Machine with 10000 training images (HAM10000) data set, which consists of 10015 dermoscopic images collected from different populations.

**Preliminaries**

**Convolutional Neural Network (CNN)**. CNN is a popular deep neural network model whose architecture design is inspired by the biological fact that the neuron connectivity patterns resemble the anima visual cortex, and for specific stimuli only the cortical neurons in a small region (a.k.a. receptive field) will respond. CNN has been shown to be very effective on computer vision tasks such as image and video analysis. Two basic operations in CNN are convolution and pooling. Convolution convolves the whole image with a small filter mimicking the response of the receptive field (defined by the filter) centered around every pixel, and the resultant map after convolution will further go through a nonlinear activation function to obtain the final response. The pooling layer is for downsizing the image maps. A typical CNN architecture includes a series of convolution and pooling layers. The final resultant feature map will be stretched into a vector and fed to a fully connected layer for endpoint task (such as classification).

**Generative Adversarial Network (GAN)**. GAN is a class of machine learning models that generates by mimicking real data distribution. There are two networks in the GAN model: a generative network that generates the data and a discriminative network to evaluate the generated data. The goal of GAN is to generate data objects as real as possible through zero-sum gaming. Many variants of GAN have been proposed in recent years. Particularly, in our case, we want to leverage the GAN technique for generating more skin image samples, and those samples should be not only real, but also can help improve the lesion classification performance. Mirza and Osindero proposed conditional GAN which makes both the data generation and discrimination probability conditioned on the data classes. Odena et al. further proposed “Auxiliary Classifier GANs” (AC-GAN) by expanding the conditional GAN. In the AC-GAN framework, the generator accepts a class label or condition along with the noise, while the discriminator estimates the probability that the image is real or fake along with the class probability. The authors note that although this modification is not tremendously different from previous works, it provides stability to the GAN’s notoriously unstable training process, in avoidance of problems like mode collapse, where the model collapses to a single mode; and training collapse, where the model stops improving.

**Methods**

As illustrated in Figure 1, there are 4 modules in our pipeline, and we will introduce them in detail in this section.
Data Preprocessing. The HAM10000 skin lesion dataset used for this work contains about 10,000 RGB images of skin lesions at a resolution of 450x600 pixels. For the purposes of this work we used 3 of the 7 skin lesion classes represented in the dataset that provided adequate samples for our GAN to learn conditional distributions, leaving a relatively small dataset of around 8900 images belonging to the disease classes of melanoma, nevus, and benign keratosis (the excluded images belonged to the 4 other classes).

Of the selected 8900 samples, 80% were used for purposes of training the models and 20% were set aside as for a final, testing set (Table 1), with both sets stratified. While training the CNNs we set aside 20% of the training dataset for internal validation.

The training set was preprocessed by application of the Gray World color constancy algorithm followed by a random 90% square crop (for more focus on the regions of interest), while for the testing dataset we took a 90% square center crop of the images. For training the CNNs, all images were then resized to 224x224 pixels, while for training the GAN, all images were resized to 64x64 pixels (later upsampled to match the CNN input size).

<table>
<thead>
<tr>
<th></th>
<th>Melanoma</th>
<th>Nevus</th>
<th>Benign Keratosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>890</td>
<td>5364</td>
<td>879</td>
</tr>
<tr>
<td>Testing</td>
<td>223</td>
<td>1341</td>
<td>220</td>
</tr>
</tbody>
</table>

Data Augmentation. We first augment the data with traditional methods, which have been shown to reduce overfitting when the number of samples is lacking. For each training sample we applied random distortions in contrast and brightness, clipped zoom, rotations, and flips, generating 4 samples per training sample. The validation and test sets were not augmented.

In addition, we also implemented the AC-GAN model for generating additional images. We followed the framework of a two-network minimax game. The generator accepted the product of random normal noise and a class embedding, outputting an image of dimensions 64x64x3, with a tanh function bringing the pixel values into the range [-1, 1]. The discriminator accepted both real and fake images and output both the probabilities of the images being real, with a sigmoid function; and the probability of the image belonging to a specific class, with a softmax function.

In our implementation, we used convolutional transpose layers in the generator network and convolutional layers in the discriminator network. For both of them we used Leaky Rectified Linear Units (Leaky ReLU) activation functions and batch normalization layers, both of which have been found to be able to help improve the training stability in GANs. The generator and discriminator models are displayed in Figure 2.

In order to address the problem of mode collapse, where the generator only learns one or a few modes of the training data’s distribution, we implemented mini-batch discrimination layers in the discriminator network. During the training process, when the discriminator is fed both real and generated mini-batches, the mini-batch discrimination layers compute the entropy of the mini-batch. Since mini-batches with much lower entropy than real mini-batches are more likely to be generated, the discriminator can provide feedback to the generator in order to increase sample diversity.
Classification Model. In order to promote the robustness of the classification model, we adopted a model ensemble scheme, instead of just using a single model. Popular image classification architectures, including VGG-16, DenseNet, Xception, and Inception ResNet v2, are adopted as base classification models\cite{VGG16, DenseNet, Xception, InceptionResNet}. As VGG simply stacks convolutional and pooling layers it is greatly inefficient, having an unfavorable proportion of parameters to accuracy. More recent architectures employ microarchitectures to improve performance such as the Inception module, residual connections, and dense connections between layers, aiming to allow networks to be deeper and more accurate. Xception modifies the Inception v3 architecture by stacking depthwise convolutions with residual connections and demonstrated a more efficient use of model parameters. Inception ResNet v2 adds residual connections to Inception v3, aiming to accelerate the training process and perhaps improve accuracy. DenseNet, by using dense connections between layers, encourages parameter sharing and feature reuse, thus achieving higher performance with fewer parameters. The different approaches taken by the different architectures allow them to have variance in their test-time errors, allowing an ensemble to of architectures to reduce variance and improve accuracy.

Because we are handling a multi-class classification problem, the categorical cross entropy (CCE) is adopted as the loss function for training the individual deep learning classifiers. Moreover, to further consider the size of different classes, we weight the cross entropy of different classes by their empirical prior probabilities calculated by Eq.\>(2). The final objective loss is shown in Eq.\>(1), where $\hat{y}_j$ and $y_j$ are vectors representing the predicted softmax labels and true one-hot labels, respectively, with $j$ as the index of the classes and $J$ as the number of classes. $\hat{y}_j$ is computed from softmax as shown in Eq.\>(3), where $s_j$ is the input to the softmax layer for class $j$.

\[
CCE = - \sum_{j=0}^{J-1} w_j y_j \log (\hat{y}_j) \tag{1}
\]

\[
w_j = \frac{\text{total} \# \text{of samples}}{\# \text{of samples of class } i} \tag{2}
\]

\[
\hat{y}_j = \frac{e^{s_j}}{\sum_{l=0}^{J-1} e^{s_l}} \tag{3}
\]

In order to evaluate the performance of the models, we adopted balanced multiclass accuracy (BACC), defined in Eq.(4). The true positive predictions of each class are denoted by $TP$, the false positive predictions of each class are denoted by $FP$, $c$ denotes a class, and $N$ denotes the number of classes.

\[
BACC = \frac{1}{N} \sum_{c=0}^{N-1} \frac{TP_c}{TP_c + FP_c} \tag{4}
\]

For the ensemble strategy, we implemented mean ensemble, where the softmax layers of the CNNs are averaged to get the image’s predicted class labels; and an RBF-kernel support vector machine (SVM)\cite{SVM}. With the SVM ensemble method, we trained the ensemble on concatenations of softmax layers from the validation data, evaluating on the hold-out test set.

For a baseline, we compared our model to scale-invariant feature transform (SIFT) with an SVM\cite{SIFT}. In this process, the RGB images are converted to grayscale. The images are then clustered with the k-means algorithm by their SIFT keypoints (128-length vectors), in order to perform vector quantization. An RBF-kernel SVM is then fit and evaluated with the dictionary of keypoints.
**Model Interpretation.** LIME is a package built to provide model-agnostic interpretations for classifiers. Being model-agnostic means that LIME acts without knowledge of the internal workings of a classifier, treating it as a black box and locally learning a mapping of input to output.

For interpretation of our models’ decisions, LIME also treats the CNNs as black boxes, perturbing an image it feeds to a given CNN and estimating the CNN’s decision function. The CNN’s decision function is estimated through a sparse linear model around a single image. By learning this decision function, LIME is able to highlight superpixels of images that lead to certain diagnoses by CNNs based on the importance of these superpixels.

**Experiments**

For all models we started their training with the ImageNet weights (available from Keras Applications). We used the Adam optimizer with a mini-batch size of 32 and learning rate of 0.0001, training for 20 epochs with early stopping. As the models displayed a tendency to overfit the training data, training was stopped when the validation accuracy and loss no longer improved. Moreover, DenseNet-169 was trained with GAN-generated samples (upscaled to model input size with Lanczos resampling) and traditional data augmentation.

Table 2 displays the balanced accuracies for each model and ensembles of models. In the ensembles we excluded SIFT with SVM, VGG-16, Xception, and DenseNet-121 as including these models decreased the ensemble accuracy.

Our results show that, with 20% of the image set aside as independent validation, our pipeline can achieve a classification accuracy of 0.8569 over three skin lesion classes, while the traditional image classification method with Support Vector Machine (SVM) trained on Scale-Invariant Feature Transformation (SIFT) features can only get a classification accuracy of 0.5326.

Figure 3 displays the normalized confusion matrices for various models on the testing set. The confusion matrices show the accuracy in predicting each disease across each row, e.g. in the first matrix, DenseNet-169 had 82% accuracy in predicting melanoma, 81% accuracy in predicting nevus, and 72% accuracy in predicting benign keratosis. Since the different CNNs learn different mappings of input to output, they have demonstrated different strengths in predicting the three diagnoses. While the final ensemble in the third confusion matrix is composed of more models than just DenseNet-169 and DenseNet-201 (50% GAN Augmentation Level), these confusion matrices demonstrate how an ensemble can balance out individual model predictions to improve classification accuracy. For instance, DenseNet-201’s poor melanoma classification accuracy being balanced out by DenseNet-169’s melanoma classification accuracy, and likewise, DenseNet-169’s poor benign keratosis classification accuracy is balanced out by DenseNet-201’s benign keratosis classification accuracy.

![Figure 3](image.png)

**Figure 3.** (a) DenseNet-169, (b) DenseNet-201 (50%), (c) SVM Ensemble Confusion Matrices
### Table 2. Classification Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT with SVM</td>
<td>.5444</td>
<td>.5326</td>
</tr>
<tr>
<td>VGG-16</td>
<td>.7276</td>
<td>.6838</td>
</tr>
<tr>
<td>Xception</td>
<td>.7344</td>
<td>.7275</td>
</tr>
<tr>
<td>Inception v3</td>
<td>.7537</td>
<td>.7446</td>
</tr>
<tr>
<td>Inception ResNet-v2</td>
<td>.7819</td>
<td>.7839</td>
</tr>
<tr>
<td>DenseNet-121</td>
<td>.7369</td>
<td>.7367</td>
</tr>
<tr>
<td>DenseNet-169 (no GAN Augmentation)</td>
<td>.7849</td>
<td>.7821</td>
</tr>
<tr>
<td>DenseNet-169 (with 40% GAN Augmentation)</td>
<td>.8089</td>
<td>.8009</td>
</tr>
<tr>
<td>DenseNet-201 (no GAN Augmentation)</td>
<td>.8236</td>
<td>.8030</td>
</tr>
<tr>
<td>DenseNet-201 (with 50% GAN Augmentation)</td>
<td>.8126</td>
<td>.8156</td>
</tr>
<tr>
<td>Mean Ensemble (without GAN-trained CNNs)</td>
<td>.8352</td>
<td>.8130</td>
</tr>
<tr>
<td>Mean Ensemble</td>
<td>.8599</td>
<td>.8478</td>
</tr>
<tr>
<td>SVM Ensemble</td>
<td>.8577</td>
<td>.8569</td>
</tr>
</tbody>
</table>

During the training of the CNNs, we evaluated different levels of augmentation with generated images over DenseNet-169 and DenseNet-201. The results are shown in Figure 4. Each level indicates the proportion of images that were generated and added to the training set, i.e. a level of 50% indicates the number of images equivalent to 50% of the original dataset (before traditional augmentation) were generated by the AC-GAN and added to the training set. For both DenseNet-169 and DenseNet-201, we employed augmentation levels of 0% (no additional augmentation), 20%, 40%, 50%, 60%, 80%, and 100%. DenseNet-169 achieved the highest balanced accuracy of 0.8009 at a level of 40%, and DenseNet-201 achieved the highest balanced accuracy of 0.8156 at a level of 50%. This indicates that a deeper model such as DenseNet-201 has more capacity for additional features provided by generated images compared to DenseNet-169. The trend for both models indicates an optimal level of augmentation at a moderate level with a performance drop with more generated data. As the level of augmentation increases, the classification accuracies of the models return to baseline levels, which may be attributed to the imprecision in generated data causing the models to lose generalizability to the testing set.
During training of AC-GAN, we did not use an imbalanced training procedure (e.g., training the discriminator on 10 batches for every time the generator is trained). Instead, we used a balanced training procedure, training the discriminator and generator equally, with a batch size of 64, with the AC-GAN converging in about 20,000 updates.

Figure 5 displays generated images side by side with their nearest neighbors (measured by Euclidean distance) in the training dataset. While we can see that skin lesions generated by the AC-GAN are similar in size, shape, and color to skin lesions from the actual training set, the AC-GAN seems to have produced original skin lesions rather than memorizing and reproducing skin lesions from the training set. Additionally, although some samples within classes are highly similar, there is still some diversity in the generated skin lesions both within each class and between the three classes, indicating the capability of AC-GAN for capturing of multiple modes of the data distribution and the absence of mode collapse.

In order to validate the acquired representation of the AC-GAN, we used latent space interpolation, which performs linear interpolation between two noise vectors (with the same class label). The results are shown in Figure 6. As the generated samples displayed smooth transitions with an absence of “holes” in the latent space of the AC-GAN, we can conclude the AC-GAN has learned meaningful features of the training distribution, rather than overfitting and producing discrete transitions.
Figure 5. Images Generated by AC-GAN*

*Leftmost column of each class of images holds the generated images, while the 3 columns to the right of it hold the 3 nearest neighbors for each generated image.

Figure 6. Latent Space Interpolation. The leftmost and rightmost images were generated from vectors sampled from a normal distribution, and the other images were generated from interpolations between the two original vectors.

For diagnosis algorithms such as a CNN to be used in real clinical decision support, we cannot just provide the classification score, but also demonstrate the evidence on how the classification decision is made. We use a model-agnostic method, named LIME, to “open a black-box” of CNN and achieve such a goal.
Figure 7 displays a few examples of positive diagnoses by Inception ResNet-v2 explained by LIME. The region bounded by a yellow line or highlighted for each lesion contain the top groups of pixels for each prediction. While LIME appears to highlight meaningful regions of the skin lesions, it may be lacking in specificity due to the difficulty of the classification task for both machines and humans. However, LIME does show that regions of the skin lesions are considered foremost before the backgrounds of the images, validating that the CNN is using relevant information for its diagnoses. There are also inherent limitations of the LIME’s model agnostic approach that come with its wide applicability, including its inability to use information from the weight activations of the CNN.

**Figure 7.** True positive diagnoses for melanoma (0), nevus (1), and benign keratosis (2) from the test set with Inception ResNet-v2.

All CNNs and the AC-GAN were trained on an Amazon Web Service p3.2xlarge EC2 cloud instance, which has an NVIDIA Tesla V100 GPU, using the Keras deep learning framework with a TensorFlow backend. We release our source code and a demo web interface at [https://github.com/alxiang/lesion-GAN](https://github.com/alxiang/lesion-GAN).

**Conclusions and Discussions**

From Table 2, the results of the classification task, the top individual model was DenseNet-201 with GAN augmentation, with a test-time accuracy of .8156, while the less complex SIFT with SVM and VGG-16 models were unable to compete. For model ensemble strategies, averaging and SVM led to better classification accuracies, and we can also observe the expected decrease in variance obtained from combining multiple predictions.

We also found models tended to overfit without the regularization effect of data augmentation and converged faster by starting with the pre-trained ImageNet weights. We still observed overfitting with data augmentation and early stopping, which suggests that our exploitation of the data is still suboptimal, and more extensive data augmentation (e.g., with GAN-based methodologies) may improve results.

The results of the AC-GAN, displayed in Figures 5 and 6, showed that it can capture meaningful representations of the data, which correspond to the multiple modes of the data distribution in different classes. Moreover, the latent space interpolation experiments with the AC-GAN provide additional validation for the generator’s learned distribution. While generating 64x64 images with the AC-GAN was stabilized by inclusion of the class condition and batch diversity during training, we were unable to scale up the GAN to 256x256, closer to the desired input size of the model.

For the classification task, there is possible room for improvement through modifications to the CNN models including test-time augmentation, more advanced ensemble strategies, and more extensive data augmentation. Even deeper models are also promising, as they may have the useful capacity for deeper features in the training dataset.

Additionally, scaling up the resolution of images generated by the AC-GAN while preserving meaningful features of the training data could improve training of the CNNs, though in our experience, higher-resolution GANs were highly unstable to train. In future attempts of high-resolution image generation, generation prowess of different GANs can be evaluated through their aid in improving the accuracy on the classification task.

**Acknowledgements**

The work of Fei Wang is supported by NSF IIS-1750326
References

Analysis of Spatial Trends in Smoking Status Among Patients with Obstructive Airway Diseases Highlight Potential for Targeted Smoking Cessation Interventions

Sherrie Xie, BS, Rebecca A. Hubbard, PhD, and Blanca E. Himes, PhD
Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Because chronic obstructive airway diseases like asthma, chronic obstructive pulmonary disease (COPD) and asthma-COPD overlap (ACO) increase individual susceptibility to the harmful effects of cigarettes, smoking cessation programs could strengthen their public health impact by targeting smokers with these conditions. We performed spatial analyses on data derived from the Electronic Health Records (EHRs) of 25,119 asthma, 3,323 COPD, and 3,620 ACO patients and a community-based health survey of 18,740 residents to identify regions in the Greater Philadelphia Area with a high density of current smokers among patients with obstructive airway diseases and the general population. We identified areas in North and West Philadelphia with high prevalence of current smokers across all patient groups and community members that should be prioritized in smoking cessation initiatives. Neighborhood deprivation, which was linked to patient data using residential geocodes, was associated with greater smoking prevalence in these regions.

Introduction

Cigarette smoking is a leading cause of morbidity and premature death in the U.S. and has been a target of many federal and local policy initiatives.\textsuperscript{1-4} Asthma and chronic obstructive pulmonary disease (COPD) are common obstructive airway diseases that increase individual susceptibility to the harmful effects of cigarettes. Among individuals with asthma, smoking increases symptom severity, accelerates lung function deterioration, and decreases therapeutic response to corticosteroids.\textsuperscript{5-8} Most COPD-related deaths are caused by smoking, and the mortality risk for smokers with COPD is twice that of smokers with other conditions.\textsuperscript{1, 9} People who have both asthma and COPD, referred to as asthma-COPD overlap (ACO), have more comorbidities and exacerbations than those with asthma or COPD alone.\textsuperscript{10-12} While ACO had been under-studied partly because asthma and COPD trials commonly excluded patients with both conditions, the recognition that people with ACO have worse outcomes has resulted in it being a topic of recent interest.\textsuperscript{13, 14}

While no study has explicitly considered the effects of smoking cessation on people with ACO, the beneficial effects of smoking cessation on those with asthma or COPD are well documented. Smokers with asthma who quit smoking experience decreased symptoms and improved lung function,\textsuperscript{7, 15} while smokers with COPD who quit smoking experience slower lung function decline and improved survival\textsuperscript{16, 17} and smoking cessation is considered the most effective treatment for COPD.\textsuperscript{17} Given the high prevalence of obstructive airway diseases in the U.S.\textsuperscript{18, 19} and the substantial benefits that smokers with these conditions stand to gain by quitting, smoking cessation initiatives can increase their public impact by reaching these vulnerable individuals.

Here, we leveraged data derived from Electronic Health Records (EHRs) and a community-based health survey to understand how smoking cessation could geographically target patients with asthma, COPD, and ACO. Our study aimed to (1) determine how smoking cessation among patients with asthma, COPD, or ACO compares to that of a representative community sample, (2) identify factors associated with smoking status among patients vs. community members, and (3) identify geographical areas with high densities of current smokers that should be prioritized for smoking cessation interventions.

Methods

Study population. Our study population consisted of (1) adults with asthma, COPD, or ACO encountered at the University of Pennsylvania Health System (UPHS) in 2012-2016, and (2) residents of Southeastern Pennsylvania who were part of the Southeastern Pennsylvania Household Health Survey (SEPA-HHS, henceforth referred to as HHHS). UPHS operates several large hospitals, specialty medical centers, and satellite clinics in the Greater Philadelphia Area, including five counties in Southeastern Pennsylvania (Philadelphia, Bucks, Chester, Delaware, and Montgomery counties, henceforth referred to as SEPA). We obtained de-identified, patient-level data from Penn Data Store, the
Variables extracted included geocoded patient residential addresses, codified demographic information (i.e., gender, birth year, race/ethnicity, height and weight measures), as well as codified smoking history (i.e., whether a patient is a current smoker, has quit smoking, or never smoked). Medications prescribed during encounters were obtained from codified entries as well as NLP-extracted values from clinical notes. Patient insurance class was obtained from an encounter-level codified billing field that was re-categorized into four groups: Private Insurance, Medicare, Medicaid, and Other/no insurance. Patients were assigned the most frequently occurring insurance class among all encounters. Patient age was defined as age at first encounter.

For the current study, patients were included if they had: (1) at least one ICD-9 or ICD-10 code for asthma (i.e. ICD-9 493* or ICD-10 J45*) or COPD (i.e. ICD-9 491*, 492*, or 496*; ICD-10 J41*, J43*, or J44*), (2) non-missing smoking history, (3) non-missing residential geocode and (4) residence within SEPA. Patients were classified as having asthma if they had at least one ICD code for asthma, prescription for a short-acting β2-agonist, and did not meet the criteria for COPD. Patients were classified as having COPD if they had at least one ICD code for COPD, prescription for a short-acting β2-agonist or short- or long-acting muscarinic antagonist, and did not meet the criteria for asthma. Patients were classified as having ACO if they met the diagnosis and medication criteria for both asthma and COPD.

The HHS, a community-based survey of SEPA residents conducted by the Public Health Management Corporation (PHMC) in 2012 and 2015, was a random-digit-dialing mobile and home telephone survey that aimed to capture the health status and behaviors of the local population and included survey items on individual-level demographics (e.g., age, gender, race, BMI, health insurance status) and smoking history. PHMC recorded residential street addresses or nearby cross-streets of survey respondents, which were used to match respondents to residential census tracts.

Neighborhood disadvantage metrics. While little information on socioeconomic status (SES) is collected in EHRs, residential geocodes can be used to link patient data to high-resolution information on neighborhood environments. We sourced data on neighborhood SES from the Neighborhood Atlas, an online tool that provides a standardized score for neighborhood disadvantage, termed the area deprivation index (ADI), for all block groups in the U.S. ADI is a factor-based score that captures multiple dimensions of SES (poverty, education, employment, housing quality) by combining several measures from the 2013 American Community Survey Five Year Estimates. It has been validated for detecting socioeconomic gradients in a number of health outcomes, including cancer mortality and hospital readmission risk. Block group ADI is reported as a national percentile ranking (taking integer values 1-100) indicating each block group’s economic deprivation relative to all others in the U.S., with higher values representing greater disadvantage. The geographic distribution of ADI in the SEPA region is shown in Figure 1. We determined neighborhood disadvantage for each patient by using residential geocodes to link their EHR-derived data to their block group ADI, which was then transformed to an ordinal variable with four levels (1-15, 16-50, 51-84, and 85-100).

Statistical analysis. Three smoking-related rates were computed for each patient group (asthma, COPD, ACO) and HHS: current smoker prevalence, ever smoker prevalence, and smoking cessation rate. Current smoker prevalence was defined as the proportion of subjects who currently smoke, and ever smoker prevalence was defined as the proportion of subjects with a positive smoking history. Smoking cessation rate was defined as the proportion of subjects who quit smoking among those with a positive smoking history. Note that while the total population served as the denominator...
for current and ever smoker prevalence, only those with a positive smoking history formed the denominator for smoking cessation rate.

Rates for community members were calculated using survey weights provided by HHS to correct for nonresponse bias, allowing estimates to more accurately reflect the SEPA population. Patient rates were standardized by gender, race, sex, and county residence using weighted HHS as the standard population via the direct method. Briefly, patients within each disease group were cross-tabulated by gender, race/ethnicity (“Non-Hispanic black,” “Non-Hispanic white,” “other”), age (18-34, 35-54, 56-74, >75), and county residence, for a maximum of 120 cells, and cells were collapsed in some patient groups to avoid cells with count 0. Subsequently, standardized rates for each patient group were calculated as the weighted average of cell-specific rates, where weights represented the relative frequency of each demographic combination in the weighted HHS sample.

We produced choropleth maps for rates aggregated at the census tract level to visualize the geographic distribution of smoking behaviors across the three patient groups and HHS. To correct for instability of rates estimated from small samples (median census tract n = 7, IQR = 3-15), we applied an empirical Bayes estimation method for small area prevalence of non-rare conditions, where crude rates were corrected towards local, county-wide means to a degree inversely proportional to the sample size. Regional differences in patient smoking behavior compared to that of local community members were visualized by subtracting HHS rates from those of each patient group for all census tracts.

We performed two-tailed Wald tests to test the null hypothesis that the difference between each smoking-related rate for each patient group vs. HHS equaled 0. To determine whether demographic factors (i.e., gender, race, age, BMI category, insurance billing class, county residence, and neighborhood disadvantage) were associated with smoking cessation for each group, we used multivariable logistic regression analyses. A separate model was constructed for each group (asthma, COPD, ACO, and HHS). Statistical analyses were conducted with a type I error rate of 0.05 and performed in R version 3.5.1.

**Geospatial analysis.** We identified significant smoking cessation cold spots (i.e., regions where the relative density of current smokers was high) for each patient group and HHS using methods described previously. Briefly, we used generalized additive models (GAM) to estimate the log odds of smoking cessation among subjects with a lifetime smoking history as a function of their location of residence before and after adjustment for individual-level covariates. Because residential geocodes were not provided by HHS, the residential location of each HHS respondent was approximated with the mean center of population for their census tracts of residence. The GAM included a bivariate smooth term S(x,y) smoothed using a loess function, a locally-weighted regression smoother which adapts to local variation in patient/respondent density. The amount of neighboring points utilized by the loess was determined by a span size optimized by minimizing the Akaike Information Criterion.

We used GAM to determine the log odds of smoking cessation on a grid of 10,000 points arranged across SEPA, which was converted to an odds ratio by using the global odds of smoking cessation within each patient group as reference. Two GAM models were fit for each patient group and for HHS respondents: one with and one without individual-level covariate adjustment, for a total of eight models. Variables selected for adjustment were the significant independent predictors of smoking cessation as determined by multivariable logistic regression: gender, race, age at first encounter, BMI category, insurance billing class and neighborhood disadvantage. County residence was not included in adjusted models to prevent over-fitting. Global tests were performed to test the null hypothesis that smoking cessation rates did not vary by geographic location by running 999
permutations of the assignment of cases (those who quit smoking) and controls (those who did not) over all residential locations. Given a significant global test, a local test was performed to identify areas of increased or decreased odds of smoking cessation. The model fit to the permutations was used to produce a distribution of log odds at each point on the grid, and points that ranked in the upper and lower 0.5% of the permutation distributions were defined as “hot spots” and “cold spots”, respectively. GAM analyses were performed in R using the MapGAM package.

Results

Characteristics of patient groups and community members. A total of 71,029 patients encountered by UPHS between 2012 and 2016 met the criteria for asthma and/or COPD. Of these, 16,530 (23%) were excluded for missing residential geocodes, 12,655 (18%) were excluded for residence outside SEPA, and 4,782 (7%) were excluded for missing smoking history, leaving 37,062 patients for geospatial analysis and the calculation of smoking-related rates. A total 20,066 HHS respondents were surveyed in 2012 and 2015. Of these, 102 (0.5%) were excluded for missing smoking history, 555 (3%) were excluded for missing residential census tract information, and 669 (3%) were excluded for missing demographic information (age or race/ethnicity), leaving 18,740 respondents for estimation of local population rates.

Demographic characteristics of UPHS patients and HHS respondents are provided in Table 1. The distribution of characteristics vary considerably between patients and HHS respondents, accentuating the importance of adjusting group-specific rates by demographic factors so that they can be compared appropriately. The median

| Table 1. Characteristics of 2012-2016 UPHS patients with asthma, chronic obstructive pulmonary disease (COPD) and asthma-COPD overlap (ACO) and respondents surveyed by the 2012 & 2015 Southeastern Pennsylvania Household Health Survey (HHS). Shown are median (IQR) for age and counts (%) for other variables. |
|---------------------------------|--------|--------|--------|--------|
|                                | Asthma | COPD   | ACO    | HHS    |
| Total                          | 25,119 | 8,323  | 3,620  | 18,740 |
| Age (IQR)                      | 40 (28, 54) | 67 (59, 76) | 61 (53, 71) | 53 (43, 64) |
| Female                         | 18,041 (71.8) | 4,181 (50.2) | 2,380 (65.7) | 11,638 (62.1) |
| Race                           |        |        |        |        |
| Non-Hispanic White             | 11,654 (46.4) | 5,073 (61.0) | 1,446 (39.9) | 13,095 (69.9) |
| Non-Hispanic Black             | 10,178 (40.5) | 2,645 (31.8) | 1,924 (53.1) | 3,900 (20.8) |
| Hispanic                       | 1,201 (4.8) | 117 (1.4) | 101 (2.8) | 936 (5.0) |
| Other                          | 2,086 (8.3) | 488 (5.9) | 149 (4.1) | 809 (4.3) |
| County residence               |        |        |        |        |
| Philadelphia                   | 13,721 (54.6) | 4,427 (53.2) | 2,425 (67.0) | 7,260 (38.7) |
| Bucks                          | 1,646 (6.6) | 685 (8.2) | 179 (4.9) | 2,743 (14.6) |
| Chester                        | 3,700 (14.7) | 1,209 (14.5) | 362 (10.0) | 2,758 (14.7) |
| Delaware                       | 3,367 (13.4) | 1,061 (12.7) | 363 (10.0) | 2,821 (15.1) |
| Montgomery                     | 2,685 (10.7) | 941 (11.3) | 291 (8.0) | 3,158 (16.9) |
| Smoking status                 |        |        |        |        |
| Never                          | 15,773 (62.8) | 809 (8.1) | 707 (17.9) | 10,432 (55.7) |
| Quit                           | 6,007 (23.9) | 5,007 (50.2) | 1,885 (47.7) | 5,252 (28.0) |
| Current                        | 15,773 (13.3) | 809 (25.1) | 707 (26.0) | 3,056 (16.3) |

Figure 3. Standardized current smoker prevalence, ever smoker prevalence, and smoking cessation rate colored by patient group or HHS. Error bars represent 95% confidence intervals.
ages of COPD and ACO patients were older than that of HHS respondents, while the median age of asthma patients was younger. Asthma and ACO patients were disproportionately female (asthma 71.8%, ACO 65.7% versus COPD 50.2%) and Non-Hispanic black (asthma 40.5%, ACO 53.1% versus COPD 31.8%), consistent with findings from previous studies of asthma and COPD in UPHS and other patient populations. All groups were enriched for subjects that lived in Philadelphia relative to other counties, though this enrichment was more pronounced in patient groups than HHS.

**Differences in smoker characteristics between patient groups and community members.** Smoking history varies markedly by age among asthma, COPD and ACO patients, and HHS respondents (Figure 2). Although COPD and ACO patients have higher proportions of current and ever smokers for all age groups, the distributions of ever smokers show a roughly unimodal distribution across all groups, with highest rates in middle to older age (45-74). The distribution of current smokers is also roughly unimodal across all groups, with peaks that occur at relatively younger ages (25-54) compared to ever smoker distributions.

Standardized rates of smoking and smoking cessation vary by patient group (Figure 3). Current smoker prevalence was higher among COPD (35.5%) and ACO patients (31.5%) compared to HHS respondents (17.9%), (p < 0.001 for both comparisons). However, current smoker prevalence was lower among asthma patients (10.3%, p < 0.001). Similarly, ever smoker prevalence was higher in COPD and ACO patients compared to HHS respondents (COPD = 77.7% and ACO = 71.9% vs. HHS = 43.4%, p < 0.001 for both comparisons), while ever smoker prevalence was

![Choropleth maps of SEPA census tracts shaded according to subject counts (top row) or smoking-related rates (bottom three rows). Rates were corrected for small sample variability via an empirical Bayes method in which crude estimates were corrected toward county means to a degree inversely proportional to tract-level sample size. Tracts with no data to calculate rates are colored white, and county boundaries are represented by black lines.](image)

**Figure 4.** Choropleth maps of SEPA census tracts shaded according to subject counts (top row) or smoking-related rates (bottom three rows). Rates were corrected for small sample variability via an empirical Bayes method in which crude estimates were corrected toward county means to a degree inversely proportional to tract-level sample size. Tracts with no data to calculate rates are colored white, and county boundaries are represented by black lines.  

1260
lower among asthma patients (38.1%, p < 0.001). Rates of smoking cessation were similar between ACO patients and HHS respondents (57.8% vs. 58.7%, p = 0.47). Smoking cessation rates were also similar between COPD patients and HHS respondents, though the difference reached marginal significance (55.4% vs. 58.7%, p = 0.03). In contrast, smoking cessation rates among asthma patients were significantly higher than that of HHS respondents (74.0% vs. 58.7%, p < 0.001).

**Geographic distribution of chronic airway disease patients by smoking status.** Current and ever smoker prevalence estimates for Philadelphia census tracts were generally higher than those of the surrounding counties, and this trend held across patient groups and HHS (Figure 4). In addition, smoking cessation rates were generally lower in Philadelphia census tracts than those of the surrounding counties across all groups. Among the suburban counties, smoking rates appeared to be higher in Chester and Delaware compared to Bucks and Montgomery, especially among COPD patients, though this trend was less striking among HHS respondents. As a result, positive rate differences were evident in Chester and Delaware counties, where smoking rates tended to be higher in patient groups compared to HHS (Figure 5), though we did not test for statistical significance. The higher smoking rates among COPD and ACO groups compared to HHS extended across all SEPA census tracts, and this difference in rates does not exhibit a discernible spatial trend (Figure 5).

**Neighborhood disadvantage measures.** Block group-level neighborhood disadvantage measures are shown in Figure 1 for the five SEPA counties studied. Philadelphia county encompasses many highly disadvantaged neighborhoods, with over one-fifth of its block groups ranking among the nation’s most disadvantaged (91-100th percentile). The surrounding suburbs (Bucks, Chester, Montgomery, Delaware counties) are markedly more affluent, with large regions that rank among the nation’s bottom 10th percentile in neighborhood disadvantage.

**Factors associated with smoking cessation.** Multivariable analyses found that older age and overweight/obese BMI were consistent independent predictors of smoking cessation across patient groups and HHS (Table 2). On the other hand, Medicaid insurance (a crude indicator of low individual-level SES) and neighborhood disadvantage (an indicator of low neighborhood-level SES) were consistent independent predictors of subjects not quitting smoking.

**Spatial trends in smoking cessation rates.** Smoking cessation rates were heterogeneous in SEPA across all groups (Figure 6). The global test statistic for spatial heterogeneity was significant for all 8 GAMs (p<0.01), indicating there was significant spatial heterogeneity in smoking cessation rates, and these trends were not fully explained by the covariates considered. The cold spots identified by the crude models (top row of Figure 6) showed considerable overlap, particularly in North and West Philadelphia, suggesting that smoking cessation rates were low (i.e., relative proportion of current smokers was high) for all patient groups and community members in these regions. While spatial trends were still present in the adjusted models (bottom row of Figure 6), the cold spots in North and West Philadelphia were absent, suggesting that the decreased smoking cessation rates in these regions were explained by spatial variation of the covariates considered.

**Discussion**

A strong evidence base supports the effectiveness of community-based interventions for smoking cessation. Policy makers interested in maximizing the impact of smoking cessation programs determine how to target groups that are most susceptible to the harmful effects of cigarettes and identify regions with the greatest number of current smokers.
We leveraged data derived from EHRs and a community-based survey to understand the smoking characteristics of patients with asthma, COPD and ACO relative to those of local populations, and we determined where smokers in these groups live. We found smoking rates among ACO and COPD patients to be nearly twice as high as those estimated from community members, and smoking rates among asthma patients to be somewhat lower than local community estimates. In addition, while smoking cessation rates were not substantially different between COPD and ACO patients vs. community members, cessation rates were significantly higher among patients with asthma. The high prevalence of smoking among patients with COPD and ACO highlights the need for smoking cessation programs that target these patients. Although smoking rates were lower among patients with asthma, the number of smokers with asthma were substantial, accounting for over 90% of smokers in our patient population. Because asthma is highly prevalent in Philadelphia,35 smokers with asthma should not escape the notice of local policy makers.

Table 2. Factors associated with smoking cessation among UPHS patients with a positive smoking history, stratified by patient group and HHS. Adjusted odds ratios (ORs) were derived from multivariable logistic regression models with smoking cessation as the outcome. Shown are ORs and 95% confidence intervals, with significant results in bold.
Factors associated with smoking cessation were similar across patient groups and HHS. Consistent with reports from the U.S. and other developed nations that the most deprived groups have the highest smoking rates and the lowest rates of quitting,36-38 those with a positive smoking history across all groups were less likely to quit if they had low individual-level SES (coarsely identified by Medicaid insurance status) or lived in disadvantaged neighborhoods. These results reaffirm the need to target smoking cessation interventions toward disadvantaged communities. Obesity also had a robust association with subjects quitting smoking, though this was likely an instance of reverse causation, as smoking suppresses appetite and individuals who quit smoking tend to gain weight.39, 40

Our unadjusted geospatial models identified consistent cold spots for smoking cessation in North and West Philadelphia (Figure 6 top row), regions with high neighborhood disadvantaged relative to SEPA and nationally (Figure 1). The loss of these cold spots in adjusted models confirmed the association of smokers’ neighborhood socioeconomic environment with quitting smoking status. Additionally, our results are promising from a policy perspective because the consistency of these cold spots indicates that initiatives that target areas with low smoking cessation rates in the general population may also target similar high-needs areas for populations with specific diseases. However, overlapping hotspots in Bucks and Montgomery county remained in the adjusted models for asthma and COPD patients, and a partially overlapping hotspot remained in the adjusted model for HHS. These hotspots suggest that the higher rates of smoking cessation in these regions cannot be fully explained by their relative affluence or the spatial distribution of the other covariates considered. Investigating the smoking-related policies in these communities could uncover the drivers of high smoking cessation rates in these regions and provide insights on effective steps that can be pursued in regions with low cessation rates.

Our findings are subject to limitations, including those inherent to analyses of EHR-derived data. Phenotyping errors can result from the use of ICD codes for the classification of disease status, though the combination of diagnosis codes with additional EHR components, such as medication information (e.g. prescriptions of short-acting ß2-agonist or short- or long-acting muscarinic antagonist used in the present analysis), can improve phenotyping performance.41

Future analyses will incorporate other EHR data, such as information from clinical notes, to further improve phenotyping accuracy. Another limitation common to EHR-derived data is data missingness. The variable with the highest degree of missingness in our EHR-derived data was patient residential geocodes, which were missing for 23% of patients. While the distributions of basic demographic characteristics (age, race, gender) were similar between patients with and without residential geocode information (data not shown), the lack of geocode information for patients in the latter group prevented us from ascertaining information about their neighborhood SES. Thus, we could not determine whether neighborhood socioeconomic environments differed between patients with missing and available geocode information. Given the strong associations between neighborhood disadvantage and smoking

![Figure 6](image-url). The spatial odds ratio for smoking cessation calculated for asthma, COPD, and ACO patients and HHS respondents who had a positive smoking history. The top row represents odds ratio surfaces estimated from unadjusted models, while plots in the bottom row include adjustment for individual-level covariates (i.e., gender, race/ethnicity, age, BMI, insurance billing class, and neighborhood disadvantage). Regions shaded blue represent areas where the odds of quitting smoking are below the global mean and areas shaded red represent areas where the odds of quitting smoking are greater than the global mean. County boundaries are represented by grey lines. Significant cold spots (with decreased rates of smoking cessation, p < 0.005) are indicated by blue contour lines, and significant hot spots (with increased rates of smoking cessation, p < 0.005) are indicated by red contour lines.
behaviors noted in the present analysis, it is unclear how the exclusion of patients without geocodes may have biased our results. A smaller, though not insignificant, proportion of patients (~7%) were excluded for missing smoking history. A disproportionate percentage of patients who were excluded for missing smoking history were Non-Hispanic black (41.3% vs. 30.1%) and lived in more disadvantaged neighborhoods (median ADI = 69 vs. 36), indicating a systematic bias in the ascertainment of smoking history among UPHS providers. Though the relatively small number of patients excluded for missing smoking history is unlikely to change our primary results, this bias in the ascertainment of smoking history warrants further investigation. The GAM models used in our analyses mapped the spatial distribution of smoking cessation rates without explicitly modeling spatial autocorrelation, and future analyses can incorporate geographically weighted regression or spatial autoregressive models to account for spatial dependence directly. In addition, future analyses can incorporate patients with non-respiratory health conditions, such as coronary heart disease and hypertension, to understand smoking characteristics across other susceptible patient populations.

A commonly cited limitation of EHR-derived data is a lack of variables, such as socioeconomic factors, that are relevant to the study of many health outcomes. We leveraged residential geocodes available in EHR-derived data to link patient data to high-resolution data on neighborhood disadvantage. The robust associations that we found between patients’ neighborhood disadvantage and their smoking status mirrored similar trends found using data from a population-based survey, thus demonstrating the utility of augmenting EHR-derived data with external geospatial variables through the use of residential geocodes. Factors that would be of interest to studies of other health-related behaviors that can be linked to EHR-derived data via similar methods include data on green spaces, walkability, crime, and locations of tobacco or alcohol outlets.

Conclusion

The efficacy of smoking cessation programs can be greatly enhanced by targeting regions with high prevalence of current smokers, particularly among those with obstructive airway diseases, which increase individual susceptibility to the harmful effects of cigarettes. We leveraged EHR-derived data on asthma, COPD, and ACO patients and data from a community-based health survey to identify several coinciding regions in North and West Philadelphia that had high densities of current smokers across patient groups and community members. These areas should be prioritized in smoking cessation initiatives.

Acknowledgements

We would like to thank Sunil Thomas, M.B.A., T.M. and Yuliya Borovskiy, M.S. from the University of Pennsylvania Penn Data Store for extracting the data used for this project. This work was partially supported by National Institutes of Health (NIH) F31 HL142153, R01 HL133433, R01 HL141992, and P30ES013508.

References

Comparison of Smart Contract Blockchains for Healthcare Applications

Hongru Yu¹*, Haiyang Sun, BE²*, Danyi Wu¹*, and Tsung-Ting Kuo, PhD¹

¹ University of California San Diego, La Jolla, CA, USA;
² Syracuse University, Syracuse, NY, USA;
* Contributed Equally

Abstract

Blockchain and smart contracts (i.e., computer code that can be run on blockchain) are increasingly popular for healthcare applications. However, only very few implementations exist because of the complexity of the technologies. Although there are tutorials and reviews to introduce blockchain and smart contracts, a pragmatic comparison of such platforms is needed. In this study, we addressed practical considerations while building a healthcare blockchain and smart contract system, by (1) comparing technical features of platforms, (2) selecting three platforms, (3) constructing blockchain networks, (4) testing the blockchains, and (5) summarizing the experience and time used for implementation by students. We evaluated Ethereum, Hyperledger Fabric, and MultiChain, and confirmed that the selection of a proper platform depends on the requirements of the application. The findings of our study can accelerate the process and reduce the risk of adopting blockchain technology in biomedical and healthcare domain.

Keywords: Data Sharing, Interoperability and Health Information Exchange, Privacy and Security

Learning Objective: Understand smart contract blockchain platforms and their comparison for healthcare.

Introduction

Since Bitcoin,¹,² blockchain technology has evolved and is now considered as a new form of a distributed ledger, since arbitrary data can be stored in the metadata of the transactions.³ With this capability, many blockchain-based healthcare applications are emerging rapidly. Example applications include unalterable patient records,⁴ improved care data sharing/analysis without ceding control,⁵⁻⁷ and enhanced robustness for counterfeit drug prevention/detection systems in pharmaceutical supply chains.⁸

Figure 1. Example of smart contract for data queries such as “how many patients have hemoglobin test results for postmenopausal females”. (a) Traditional process of a researcher requesting a data query within a clinical data research network (i.e., from University A to University B), including the use of an electronic form as well as the need for a data concierge person to manage and respond to the queries. (b) Blockchain-based smart contract approach to manage/record the data queries automatically without human. Note that the choice of the query depends on the Institutional Review Board (IRB) approval as well as the Data Usage Agreement (DUA) signed between universities.
Among the new developments of the blockchain technology, smart contract is one of the most important advancements.9-19 Smart contracts refer to computer programs/code devised to manage smart properties, which are the digital properties recorded by blockchain.11 That is, a smart contract is a piece of computer code that can be run within a blockchain transaction. For example, a smart contract has been proposed to automate the workflow of medical licensure across multiple states.20,21 It removes the requirement for human confirmation of precursors and results, and instead relies on machines to automatically execute the logic stated in the program to accept or reject a request. Smart contracts are also transparent and immutable. An example is shown in Figure 1, demonstrating the potential to manage the data queries requested by an institution to another within a clinical data research network. Therefore, we consider smart contract as a very useful and desired feature for healthcare applications.

Many researchers have proposed to adopt blockchain and smart contracts in healthcare, genomic, or biomedical applications.4-6, 8, 22-35 but only very few have actually implemented these technologies due to their complexity.3 Therefore, understanding practical considerations while exploring smart contract technology is imperative, and can facilitate the adoption of the new technology for a wild spectrum of healthcare applications. Although there are broad tutorials and reviews about blockchain and smart contracts,36-39 a pragmatic comparison and evaluation of the smart contract platforms is needed.

Table 1. Technical features to compare smart contract blockchain platforms.

<table>
<thead>
<tr>
<th>Technical Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Smart Contract</td>
<td>The type of smart contract supported in the platform, can be either first</td>
</tr>
<tr>
<td></td>
<td>generation (i.e., supports scripting functions without full programming</td>
</tr>
<tr>
<td></td>
<td>capacity such as looping) or second generation (i.e., supports any</td>
</tr>
<tr>
<td></td>
<td>algorithm).</td>
</tr>
<tr>
<td>Programming Language</td>
<td>The language for smart contract (e.g., Solidity40-42 for Ethereum11, 43).</td>
</tr>
<tr>
<td>Network Permission</td>
<td>The supported permission type of blockchain network, can be permissionless</td>
</tr>
<tr>
<td></td>
<td>(i.e., any machine can join the network freely, as for Bitcoin Blockchain),</td>
</tr>
<tr>
<td></td>
<td>permissioned (i.e., only an authenticated machine can join the network), or</td>
</tr>
<tr>
<td></td>
<td>both.</td>
</tr>
<tr>
<td>Open Source License</td>
<td>The license to distribute derivative products (e.g., MIT44 or GPL45)</td>
</tr>
<tr>
<td>Applications</td>
<td>Proposed applications in healthcare.</td>
</tr>
</tbody>
</table>

Methods

In this study, we addressed the practical considerations while building a healthcare blockchain and smart contract system. Specifically, we compared the technical features of the smart contract blockchain platforms, selected platforms for our study, constructed a blockchain network for each of the platform, tested the blockchains, and summarized our experience and time used for the implementation by students. First, we selected technical features of the smart contract blockchain platforms based on our recent review/report.3, 38 as shown in Table 1. Based on the technical features described in Table 1, we identified three platforms to be compared in this study, out of ten popular platforms:3, 38

- **Ethereum**11, 43 is one of the most well-known decentralized smart contract blockchain platforms.11, 17, 43, 46 Ethereum is also a cryptocurrency itself, and has a large market cap.47 Microsoft adopted Ethereum as the core of its Blockchain-as-a-Service (BaaS) on the Azure cloud computing environment,48 however Ethereum can be run in any cloud computing environment. Ethereum has been used in healthcare applications and, most importantly, supports second-generation smart contracts.3 Ethereum can work with both external owned accounts and internal contract accounts.49 It also uses the concept of ‘gas’ to regulate computationally intensive smart contract code.46 Ethereum has a build-in currency (Ether), and Wei is the smallest fraction of Ether.50 The main reasons to select Ethereum were the capability to be configured as both a permission-less and a permissioned blockchain network, as well as the community-based development of the platform.
Hyperledger Fabric\textsuperscript{51-53} is an enterprise-grade cross-industry platform supported by IBM.\textsuperscript{53, 54} Hyperledger, a project under the Linux Foundation,\textsuperscript{51, 52} is a platform with various “dialects”, and we selected Hyperledger Fabric because it is one of the popular smart contract platform that requires no specific hardware to be executed.\textsuperscript{3, 38} Hyperledger also has a Composer component that supports a purpose-built modeling language for defining a business network, multiple ways of integrating blockchains with other services, and a platform for developing, testing and deploying custom business logic.\textsuperscript{25} The Hyperledger Composer, along with its Playground feature, enables developers to use pre-existing sample network frameworks to plan their network construction through a user-friendly, easily accessible web interface. There are also Software Development Kits (SDKs) available on Hyperledger Fabric ‘s website for potential C and Java developers. The main considerations to choose Hyperledger Fabric were the capability to be configured as a permissioned blockchain network and the access control feature with high granularity.

MultiChain\textsuperscript{56, 57} is a platform that currently only supports first-generation smart contracts as well as limited programs, or “smart filters”\textsuperscript{,58} It also does not yet support complex syntaxes based on Turing-complete programming languages that can execute any algorithm. Currently, a MultiChain smart filter is restricted to programming instructions that do not allow for operations such as loops and recursion. However, MultiChain provides many built-in functions to support basic building blocks for applications,\textsuperscript{58} and can be used directly for operations like sending or receiving transactions. MultiChain also supports data streams (i.e., key-value pairs) based on transaction metadata.\textsuperscript{59} Many MultiChain APIs have been developed for mainstream programming languages such as Java and JavaScript, enabling developers to run complex programs on MultiChain. The main reasons to include MultiChain were the simple setup process and the features inherited from Bitcoin Blockchain.\textsuperscript{1, 2}

Our comparative analysis was conducted by three undergraduate/graduate students with no prior experience in blockchain technologies to construct, test, and summarize the experiences related to the three selected platforms in parallel. While only one student developed code for each one of the platforms, they compared notes and discussed solutions as a group, with facilitation from a faculty member (T-TK). First, we constructed three selected smart contract blockchain networks. We used six virtual machines (VMs), each has 2-Core CPU, 8GB RAMs and 100GB storage, with 64-bit Ubuntu 14.04 operating system, on the Google Cloud Platform.\textsuperscript{60} We then tested the blockchain platforms by submitting transactions or deploying/executing smart contracts on them. Finally, we summarized our implementation experiences for these blockchain platforms, with an estimation of the required time for setup and learning time for an undergraduate or a graduate student.

**Results**

We compared the platforms by extracting the technical features,\textsuperscript{3, 38} and the results are shown in Table 2. Our experience and setup/learning time of constructing three blockchain platforms are summarized below.

- **Ethereum.** We constructed a blockchain network using Go-Ethereum.\textsuperscript{61} Then, we sent transactions with specified data and checked the transaction input for both servers were the same. Finally, we wrote, compiled, deployed and executed smart contracts using Solidity\textsuperscript{60, 41} on our blockchain network. An example smart contract pseudo code written in Solidity and executed on Ethereum\textsuperscript{11, 43} is shown in Figure 2, which corresponds to the use case described in Figure 1. The whole process to build the blockchain network took about 30 minutes, while the learning time for an undergraduate student (DW) was about 7 days.

- **Hyperledger Fabric.** We built a blockchain network using the Hyperledger Fabric images on Docker.\textsuperscript{62} Then, we send transactions to verify the connectivity of the network, and implemented/executed smart contracts using Chaincode in the Go language. An example pseudo code, corresponding to Figure 1 (b), is shown in Figure 3. The whole process to build the blockchain network took about 45 minutes, and the overall learning period for an undergraduate student (HY) was about 14 days.
**MultiChain.** We setup a blockchain network using MultiChain.\(^{57}\) Then, we sent transactions between nodes to ensure the validity of the network. We also tried to create data streams to disseminate key-value paired data. Finally, we investigated smart filters, and an example pseudo code is illustrated in Figure 4. The whole process to build the blockchain network took about 30 minutes, while the overall learning period for a graduate student (HS) was about 2 days.

### Table 2. Comparison of the technical features of the platforms and their applications.\(^3, 3^8\)

<table>
<thead>
<tr>
<th>Platform</th>
<th>Ethereum</th>
<th>Hyperledger Fabric</th>
<th>MultiChain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Smart Contract</td>
<td>Second Generation</td>
<td>Second Generation</td>
<td>First Generation</td>
</tr>
<tr>
<td>Programming Language</td>
<td>Solidity, Serpent, and LLL(^{63})</td>
<td>Go and Node.js for Chaincode(^{52})</td>
<td>JavaScript for Smart Filter(^{64})</td>
</tr>
<tr>
<td>Network Permission</td>
<td>Permissioned/Permissionless</td>
<td>Permissioned</td>
<td>Permissioned</td>
</tr>
<tr>
<td>Open Source License</td>
<td>Go-Ethereum: LGPL v3.0(^{65})</td>
<td>Apache License v2.0(^{69})</td>
<td>GPL v3.0(^{70})</td>
</tr>
<tr>
<td></td>
<td>CPP-Ethereum: GPL v3.0(^{65})</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Py-Ethereum: MIT License(^{66})</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EthereumJ: GPL v3.0(^{67}) and Parity; GPL v3.0(^{58})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applications</td>
<td>MedRec,(^{31, 32}) Patientory,(^{71, 72}) Nebula Genomics,(^{73, 74}) clinical data sharing,(^{75}) and patient monitoring.(^76)</td>
<td>Clinical data sharing,(^77) Institutional Review Boards (IRB) regulation enforcement,(^78) mobile healthcare,(^80) and medical data storage/access.(^80)</td>
<td>Privacy-preserving predictive modeling(^6, 7) and its application in anesthesia.(^6)</td>
</tr>
</tbody>
</table>

```solidity
pragma solidity ^0.4.22;
contract QuerySmartContract {
  address public ownerAddress;
  string public patientOutcome;
  string public patientAge;
  string public patientSex;
  constructor () public {
    ownerAddress = msg.sender;
  }
  function sendQuery(string _patientOutcome, string _patientAge, string _patientSex) public returns (bool) {
    patientOutcome = _patientOutcome;
    patientAge = _patientAge;
    patientSex = _patientSex;
    return true;
  }
  function getQuery() public view returns (address _ownerAddress, string _patientOutcome, string _patientAge, string _patientSex) {
    _ownerAddress = ownerAddress;
    _patientOutcome = patientOutcome;
    _patientAge = patientAge;
    _patientSex = patientSex;
  }
}
```

> s.sendQuery("hemoglobin", "postmenopausal", "female")
> "0xb3b7b73e82d517151369889768dc3a016b2d86e90472cf1d5361451e659a19c6"

> s.getQuery()
> ["0xd5a6c396a3e66b7e76949895f86dc878477f44b", "", "", ""]

> s.getQuery()
> ["0xd5a6c396a3e66b7e76949895f86dc878477f44b", "hemoglobin", "postmenopausal", "female"]

**Figure 2.** Example Ethereum Solidity pseudo code involved in a simple contract between two institutions (as described in Figure 1 (b)). University B deployed this `QuerySmartContract`, with two functions: `sendQuery` and `getQuery`, to the blockchain network. University A executed this smart contract and queried the data for “how many patients have hemoglobin test results for postmenopausal females” (the upper-right box). Then, University A retrieved the query from the blockchain, and can see the query criteria as well as the address of the contract owner (which is University B). Note that it takes some time for the smart contract to be enclosed in a verified transaction of a block, therefore the first query only returns empty values for the query. Also, this smart contract, as well as any function calls that change the state of the variables in the smart contract, will be logged in the immutable blockchain, which enables future attribution, auditing, dispute and resolution for both (and even a third-party) institutions.
Figure 3. Example Hyperledger Fabric pseudo code for the example shown in Figure 1 (b), and the execution results are shown in the upper-right box. Similar to the code shown in Figure 2, University B deployed this contract, and University A send the query (by calling sendQuery and getQuery functions).

```go
func (t *SimpleChaincode) sendQuery(stub shim.ChaincodeStubInterface) pb.Response { 
    fmt.Println("sendQuery")
    _, args := stub.GetFunctionAndParameters()
    var O, A, S string
    var patientOutcome, patientAge, patientSex string
    O = args[0]
    patientOutcome = args[1]
    A = args[2]
    patientAge = args[3]
    S = args[4]
    patientSex = args[5]
    fmt.Printf("%s, %s, %s/n", patientOutcome, patientAge, patientSex)
    stub.PutState(O, []byte(patientOutcome))
    stub.PutState(A, []byte(patientAge))
    stub.PutState(S, []byte(patientSex))
    return shim.Success(nil)
}

func (t *SimpleChaincode) getQuery() pb.Response { 
    var O, A, S string
    var err error
    O = "patientOutcome"
    A = "patientAge"
    S = "patientSex"
    outcome, err := stub.GetState(O)
    age, err := stub.GetState(A)
    sex, err := stub.GetState(S)
    if err != nil { 
        fmt.Println(err)
        return shim.Error(err)
    }
    fmt.Printf("%s, %s, %s/n", outcome, age, sex)
    return shim.Success(outcome)
}
```

University A:

```
$ multi-chain-cli query-chain
Interactive mode

> sendQuery (“patientOutcome”, “hemoglobin”, “patientAge”, “postmenopausal”, “patientSex”, “female”) > getQuery()
```

University B:

```
$ multi-chain-cli query-chain
Interactive mode

> subscribe query

> liststreamkeyitems query patientOutcome
```

Figure 4. Example MultiChain code between two universities. University A creates a stream query in the blockchain, and publishes a key (patientOutcome only, for simplicity purpose) with a value (hemoglobin, which is represented in hexadecimal number “68656d6f76c6f62696e”) in the stream. Then, University B subscribes the stream to inspect the value of the patientOutcome.
Comparison and Discussion
Based on our implementations, MultiChain is the easiest blockchain platform in terms of setup. The setup of Ethereum as a permissioned network took longer than MultiChain. Hyperledger Fabric contains more layers in its network to increase manageability and security, which can result in long setup time compared to other platforms. The installation of the software prerequisites for MultiChain and Ethereum are also easier compared to Hyperledger Fabric. On the other hand, Ethereum and Hyperledger provide the full-functional smart contract capability, while the support of smart contract for MultiChain is limited. The smart contracts are highly readable and easily programmable.

Considering the biomedical/healthcare applications, the salient characteristic of the three platforms are as follows. Ethereum is supported and maintained by a large community of developers around the world. This open property makes it a good choice while considering the long-term sustainability of the platform that the applications rely on. Hyperledger Fabric’s well-designed, multi-layered access control framework along with Hyperledger Fabric’s own certificates lead to high versatility, security, and manageability of the blockchain platform. MultiChain is designed to be a permissioned blockchain that is very simple yet powerful to use and inherits proven features of the famous Bitcoin Blockchain by forking from it.

Conclusion
To support the development of healthcare applications based on blockchain and smart contracts, we conducted a pragmatic comparison and evaluation of smart contract blockchain platforms. The main limitation of this study is that we only compared three platforms, so the scope may be restricted. Based on our results, an informatics researcher, IT expert, or technical leader in healthcare or other institution can assess the various practical aspects, such as setup/learning time and unique technical features, of the platforms. The selection of a proper platform depends on requirements of the application (e.g., maintenance for Ethereum, fine-grained access control for Hyperledger Fabric, and rapid-development for MultiChain). Such an assessment can accelerate the process and reduce the risks in adopting blockchain, an immutable, distributed, and automated technology, for biomedical and healthcare applications. We plan to identify more blockchain requirements for healthcare applications, conduct more structured and large-scale comparisons to include additional blockchain platforms (such as Corda) that may fulfill the requirements, and analyze the operation/performance aspects of the platforms in our future work.

Acknowledgements
The authors would like to thank Lucila Ohno-Machado, MD, PhD, and Jejo David Koola, MD, for very helpful discussions and support. Research reported in this publication was supported by the National Human Genome Research Institute (NHGRI), the National Library of Medicine (NLM), and the Office of the Director (OD), National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. All authors were partly funded by NIH, including T-TK (OT3OD025462, K99/R00HG009680), HS (OT3OD025462, K99/R00HG009680), HY (K99/R00HG009680), and DW (T15LM011271). This research benefited from the use of credits from the National Institutes of Health (NIH) Cloud Credits Model Pilot, a component of the NIH Big Data to Knowledge (BD2K) program. T-TK is also funded by the UCSD Academic Senate Research Grant RG084150 and Travel Award TG087935.
References


24. Culver K. Blockchain Technologies: A whitepaper discussing how the claims process can be improved. ONC/NIST Use of Blockchain for Healthcare and Research Workshop; Gaithersburg, Maryland, United States 2016.

25. Goldwater J. The Use of a Blockchain to Foster the Development of Patient-Reported Outcome Measures. ONC/NIST Use of Blockchain for Healthcare and Research Workshop; Gaithersburg, Maryland, United States 2016.


30. Ekbklow A, Azaria A, Halamka JD, Lippman A. A Case Study for Blockchain in Healthcare: “MedRec” prototype for electronic health records and medical research data. ONC/NIST Use of Blockchain for Healthcare and Research Workshop; Gaithersburg, Maryland, United States 2016.


33. Linn LA, Koo MB. Blockchain For Health Data and Its Potential Use in Health IT and Health Care Related Research. ONC/NIST Use of Blockchain for Healthcare and Research Workshop; Gaithersburg, Maryland, United States 2016.


38. Key Metrics of Blockchain Platforms.


42. The MIT License. Available from: https://opensource.org/licenses/MIT.


1273
Applying Blockchain Technology to Enhance Clinical Trial Recruitment

Yan Zhuang, BS¹, Lincoln R. Sheets, MD, PhD², Zonyin Shae, PhD³,⁴, Yin-Wu Chen, PhD³, Jeffrey J. P. Tsai, PhD³, Chi-Ren Shyu, PhD¹,²,³

1 Informatics Institute; 2 School of Medicine, University of Missouri, Columbia, MO, USA; 3 Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan; 4 Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan

Abstract

Patient recruitment for clinical trials is known to be a challenging aspect of clinical research. There are multiple competing concerns from the sponsor, patient and principal investigator’s perspectives resulting in most clinical trials not meeting recruitment requirements on time. Conducting under-enrolled clinical trials affects the power of conclusive results or causes premature trial termination. The Blockchain is a distributed ledger technology originally applied in the financial sector. Its features as a peer-to-peer system with publicly audited transactions, data security, and patient privacy are a good fit for the needs of clinical trials recruitment. The “Smart Contract” is a programmable self-executing protocol that regulates the blockchain transactions. Given current recruitment challenges, we have proposed a blockchain model containing multiple trial-based contracts for trial management and patient engagement and a master smart contract for automated subject matching, patient recruitment, and trial-based contracts management.

Introduction

Patient recruitment is essential to the success of clinical trials. Failure to meet recruitment goals in time results in a waste of funds and time, incomprehensible statistical results, and delay of the study timeline that could double the planned recruitment period¹⁻². 86% of clinical trials don’t achieve their recruitment goals on time³ and 19% of registered clinical trials were either closed or terminated due to failure to reach expected enrollment⁴. Barriers persist although there have been many research papers addressing the challenges of identifying and recruiting subjects to clinical trials over the past decades⁵⁻⁹.

Barriers to recruiting patients into clinical trials can be classified into three different categories (Table 1) based on (1) sponsor perspectives, (2) principal investigator perspectives, and (3) subject perspectives⁵⁻⁸. Sponsors initially need adequate participants for the potential trial to file an application with the Food and Drug Administration (FDA) for approval⁴. Inefficient advertising models such as radio, newspaper, physician referrals, flyers, cold calls, etc. make it difficult to meet the expectations of initial recruitment on time¹⁰. Sponsors need to design clinical trial protocols such as inclusion/exclusion criteria which can be used to check the eligibility of potential subjects. Detailed protocols can drastically narrow the subject population, which increases the difficulty of recruitment⁸,¹¹. Sponsors need to settle on trial sites without knowing the geographical distribution of future subjects⁷, but distant trial sites will deter many potential subjects⁸. For principal investigators, barriers include lack of awareness of available, appropriate clinical trials, excessive time spent to get the informed consent of participants, and insufficient trial protocols⁴,⁷,¹¹. Subject-related barriers are related to participation, such as patients’ lack of awareness of the available clinical trials; difficulty understanding complex protocols; high expenses if the trial has no clinical sites nearby; and distrust in the clinical trials⁸,¹¹,¹². With these persistent challenges, an efficient model is needed to enhance the recruitment process.

Table 1. Current recruitment barriers from different perspectives

<table>
<thead>
<tr>
<th>Sponsors</th>
<th>Principle investigators</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inefficient advertising models</td>
<td>1. Lack of awareness</td>
<td>1. Lack of awareness</td>
</tr>
<tr>
<td>3. Beforehand trial sites selection</td>
<td>3. Insufficient trial protocols</td>
<td>3. Inaccessible clinical sites</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. Distrust of clinical trials</td>
</tr>
</tbody>
</table>

Blockchain is a distributed ledger technology first applied in the financial sector¹³. The success of the Bitcoin cryptocurrency, which is one of the blockchain’s most popular applications, shows the robustness, security and consensus mechanism of the blockchain system¹⁴. Blockchain also has other features such as decentralization,
immutability, ensuring data provenance, and public auditability\textsuperscript{14}. All the transactions that have occurred in the blockchain are distributively stored into each node which can be any active electronic device without hardware requirements inside the blockchain system. Any transaction needs to be validated by the users in the blockchain before it is written into the system\textsuperscript{15}. Since the system is fully decentralized, all the transactions can be audited publicly by all the users. The public auditability feature can solve the issue of lack of awareness.

The Smart Contract is a self-executing, coded protocol agreed between senders and receivers initially proposed by Nick Szabo in 1996 to regulate all transactions on the distributed ledger system\textsuperscript{16}. Most blockchain systems have added a smart contract function to their protocols such as Ethereum and Hyperledger\textsuperscript{17}. Since the Ethereum blockchain comes with a built-in Turing complete programming language used for the smart contract, any computational problem such as patient matching for recruitment, or checking the validity of a clinical trial can be coded as a smart contract\textsuperscript{18}. Once a smart contract is deployed into the system, the system will generate an application binary interface (ABI) of the smart contract and smart contract address for users to call the smart contract functions\textsuperscript{14}. All the users need to follow the smart contract’s regulation to make a transaction. The ABI and address are viewable to all the users rather than the source code or the data stored inside the smart contract, but the functions may not be executable to any user. This depends on the user’s privilege. For example, in the clinical trial recruitment setting, the inclusion/exclusion protocol can only be input into a smart contract by the authority who is in charge of all clinical trials. This ideally would be the FDA. Only the sponsors can send a matching request to the smart contract, then the blockchain system can automatically match the potential subjects. This matching of sponsors and subjects can help the sponsor to target eligible patients, save time for principal investigators and patients to understand the protocols, and ensure the completeness of the protocol. Once the eligible subjects are identified, the sponsor can select the trial sites based on the subjects’ geographical distribution which could solve the issue of inaccessible clinical sites for subjects. The smart contract can also ensure the validity of a clinical trial by checking the sponsor’s identity and whether the trial has a National Clinical Trial (NCT) identifier number which means the study has been approved by the authority\textsuperscript{19}. Patients can rest assured that the trial is legal, and that the sponsors are reliable. Original blockchain features and smart contract functionality can solve most of the barriers for clinical trial recruitment.

Despite the features of blockchain fitting most healthcare applications, there are several common challenges of the most current blockchain models for healthcare applications: (1) inadequate public/private key management systems\textsuperscript{20}; inability to retrieve any information once a user has lost the private key, (2) loss of privacy caused by transparency of the distributed system\textsuperscript{21, 22}; ability of all the users in the blockchain to view all the data stored in the blockchain, and (3) scalability constraints\textsuperscript{22, 23} considering Ethereum can handle roughly 15 transactions per second. Most blockchain applications in the healthcare area are still in the design stage and have not yet been implemented\textsuperscript{24, 25}. The abovementioned challenges are mainly caused by blockchain’s features instead of actual operation. We have implemented a blockchain system using Ethereum blockchain which is an open source platform with a built-in smart contract function to tackle the issues for clinical trial recruitment. We have tested our system by simulating clinical trial recruitment using a real dataset to test the feasibility and provide potential solutions to current common blockchain challenges. We have made the following assumptions to perform the simulation of the recruitment process using our system: (1) the authority, each clinical site and each sponsor needs to provide at least one node in the system; (2) patients have to opt-in to our system, they authorized the system to access their health records; (3) clinical sites agreed to connect the secured Electronic Health Records (EHR) database through blockchain node; (4) all the EHR are recorded in the same standard; (5) there is a universal patient ID to map the same patient across different clinical sites; (6) the authority, each clinical site and each sponsor needs an administrator to operate the system, patients need to operate the system as well.

**Blockchain**

Blockchain is a fully distributed peer-to-peer network\textsuperscript{14}. The blockchain system runs based on users’ consent rather than being managed by a third party. To protect the users’ privacy, every user in the blockchain will have a unique key pair which contains a public key and a private key to represent their identities\textsuperscript{26}. The public key is similar to a user’s bank account. The private key is similar to a user’s signature. Every transaction needs to be digitally signed by the sender’s private key\textsuperscript{26}. Once a user makes a transaction, all the users in the blockchain can see the sender’s public key instead of their real identity. The blockchain has a built-in mechanism to check whether the sender’s public key and the signed private key match, but keeps the private key hidden from other users\textsuperscript{13}. The stability of the blockchain relies on all nodes to provide computing power to validate the transactions. Once a transaction which is one-time exchange of data is made, the transaction will be sent to each user in the blockchain for validation. If a hacker wants to hack others’ accounts to made transactions using a fake private key, the transaction will be voted down by other users and discarded. Where blockchain was originally used in the finical sector, the transaction referred to an exchange
of cryptocurrency like bitcoin. The validation process validates whether the sender has sufficient balance to make the transaction. In our system, a transaction refers to the execution of smart contract function such as authority input inclusion/exclusion criteria of a certain trial into a smart contract. All the users will validate whether the sender’s public key is the authority, the data which is criteria will be encrypted and can only be decrypted by specific users instead of all users. Blockchain is a chain of blocks. Each block contains a unique block number, last block number and all of the transactions that occurred after the last block was generated. The block numbers chain the blocks together that can be used for tracing the source of the data. For instance, the sponsors need to check the patients’ records to check the eligibility for the trial, they can check whether the records were input by the hospitals by tracing the sender of the initial transaction. Once a transaction is verified and written into the block, the record and data cannot be altered anymore.

Blockchain can be built as a “public chain” or a “private chain” meaning all the users can join the public chain without additional setups and permission. “Bitcoin” and “Ether” are applications on the public blockchain and public Ethereum blockchain, respectively. The private chain is called the “permission chain”. The creator of the private chain has control over who can join the chain. In order to join the private chain, all the nodes need to deploy the same “genesis block” which is the starting block of the private chain provided by the creator and add at least one peer node which already in the chain. The “genesis block” of a private chain determines the private chain’s characteristics such as the ability to estimate block generation rate. The genesis blocks are different from different blockchains so that they can only be acquired from the creators. In our scenario, the authorities will be the creator of the system. All the nodes provided by the clinical sites and sponsors need to get the “genesis block” file from the authority. The “genesis block” file is a JSON file that can be embedded into a program used for blockchain system installation. All the nodes also need to add the authority’s node as a peer using its IP address and node identifier which is automatically generated by the blockchain. This procedure can also be embedded into the install program provided by the authority. Since the blockchain is a fully distributed system, the creator can decide who has permission to join the chain. Other nodes have the same privileges to operate the system such as sending transactions, using smart contracts, auditing transactions, etc.

Since blockchain is not allowed to communicate with servers outside of the chain, we need to rely on a blockchain adapter to check the patient’s EHR data and push the result back to the blockchain in order to precisely match the subject with complex inclusion and exclusion criteria. For example, the sponsors from different IPs send requests to a clinical site’s adapter to identify potential trial candidates from EHR systems. Each node provided by the clinical site needs to be built as a “blockchain adapter” to join our system. Blockchain adapter is a Remote Procedure Call (RPC) server which can connect the clinical site’s own secured EHR database protected by the hospital’s firewall. RPC server executes functions after clients from different IPs send requests to the server. When a patient opts-in to the blockchain system from a clinical site, this site’s administrator needs to create a blockchain account for the patient through their node. The blockchain account will then be associated with the patient’s universal patient ID in the EHR database. The administrator needs to create accounts for clinical research coordinators (CRC) as well. The authority also needs to build its node into a blockchain adapter to communicate with the NCT database in order to check whether the NCT number and the sponsor’s information has been matched.

**Implementation**

To utilize the unique technological capability of blockchain for clinical trial recruitment, we implemented a private Ethereum blockchain system to simulate the recruitment process. The authority’s node as the creator node needs to start the blockchain system using a unique “genesis block” file, and other nodes and adapters then join the system using the blockchain identifier and IP address of the authority’s node. The whole system architecture (Figure 1) contains two modules: (1) A master smart contract is used for auto-matching of potential subjects for all trials using inclusion and exclusion criteria as shown on the left of the figure; (2) Multiple trial-based smart contracts are used for patients’ enrollments, trial management and future persistent monitoring for different clinical trials as shown on the right of the figure. In this setting, all the users can access the master smart contract so that it can reach any user in the system to perform the matching process. The trial-based contract is only available to the users to participate in that trial so that the trial-based contract cannot notice the users outside that trial. A use case is that the CRC can monitor the subject’s condition during the clinical trial through the trial-based contract, but they can’t access subjects’ records for other clinical trials in order to know who has participated in other clinical trials.
Only the selected group of users can execute specific functions in smart contracts. Users without privileges cannot see the data stored inside the smart contract (Table 2). To ensure the accuracy of input data, inclusion/exclusion criteria need to be input by the sponsors and executed by the authority, and the patients’ primary records which include demographic information, previous primary diagnosis and treatment from each visit that used for trial matching can only be input by the clinical sites. The authority has an oversight role in the system. All the clinical sites and sponsors need to get approval from the authority to provide a node to join the system. The authority will intervene in any inconsistent data such as differing patient records in the master smart contract and trial-based contract. The authority can trace the inputter of the records and investigate the reason. This setting can ensure the trial is conducted precisely under the authority’s real-time surveillance.

Module 1 requires all clinical sites to input the opted-in patient list and their primary records. Each sponsor needs to send a transaction to the authority containing the NCT identifier number and its own information for validation. After the authority validates the sponsor’s identity and the authenticity of the requested trial, the authority will input the inclusion/exclusion criteria to the master smart contract. In the meantime, a trial-based contract for this clinical trial will be generated and the contract address will be stored into the master smart contract. The sponsor can request the auto-matching process after sending the transaction to the authority. Part of the auto-matching smart contract code and the returned ABI after compiling this function is shown in Figure 2.

**Table 2.** Privileges for users to use smart contract functions

<table>
<thead>
<tr>
<th></th>
<th>Execute</th>
<th>View</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Master Smart Contract</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Opt-in list</td>
<td>Clinical sites</td>
<td>Clinical sites/ Authority/ Opted-in subjects/ Sponsor</td>
</tr>
<tr>
<td>Patients’ Information</td>
<td>Clinical sites</td>
<td>Clinical sites/ Specific patient</td>
</tr>
<tr>
<td>Inclusion/Exclusion Criteria</td>
<td>Authority</td>
<td>authority/ Specific patient</td>
</tr>
<tr>
<td>Auto-matching</td>
<td>Clinical sites</td>
<td>N/A</td>
</tr>
<tr>
<td>Trial contract info</td>
<td>Authority</td>
<td>Trial sites/ Authority/ Enrolled subjects/ Trial sponsor</td>
</tr>
<tr>
<td><strong>Trial Contract</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sponsor information</td>
<td>Authority</td>
<td>All</td>
</tr>
<tr>
<td>Patient’s trial records</td>
<td>CRC</td>
<td>Trial sites/ Authority/ Specific subjects/ Sponsor</td>
</tr>
<tr>
<td>Other functions</td>
<td>Sponsor</td>
<td>All</td>
</tr>
</tbody>
</table>

**Figure 1.** System architecture master smart contract and different clinical trial smart contracts
The patient matching is a two-step process. The first step is using the auto-matching function which can only narrow down the patient selection but cannot perform precise matching due to the complexity of the inclusion/exclusion criteria. The master smart contract will automatically match the criteria with the patients’ records stored in the smart contract to select potential subjects. Referring to the example shown in Figure 3, the master smart contract filtered out patient B due to the exclusion criteria of history of renal disease. The second step is performed by the clinical sites to precisely match the potential subjects’ health records in the hospital’s secured database with the patient’s information from the consents. The matching process is then followed by sending the result back to the blockchain through its own blockchain adapter. The exclusion criteria in the example in Figure 3 also has current tobacco use which is not recorded in patient A’s information. The master smart contract will notify patient A that there is a potential clinical trial that he/she might be eligible to participate in and need his/her authentication for the sponsor to access his/her EHR to double check with the details. The sponsor can communicate with the clinical sites which patient A visited before to check whether he/she uses tobacco currently with patient A’s consent by E-signature using the private key. The clinical site will perform precise checking for the sponsor.

Figure 2. Sample matching function and ABI ((a) The sample code of the matching function. The real function is calling several matching functions based on the criteria. (b) The ABI of the sample function which is viewable to every user.)

The patient matching is a two-step process. The first step is using the auto-matching function which can only narrow down the patient selection but cannot perform precise matching due to the complexity of the inclusion/exclusion criteria. The master smart contract will automatically match the criteria with the patients’ records stored in the smart contract to select potential subjects. Referring to the example shown in Figure 3, the master smart contract filtered out patient B due to the exclusion criteria of history of renal disease. The second step is performed by the clinical sites to precisely match the potential subjects’ health records in the hospital’s secured database with the patient’s information from the consents. The matching process is then followed by sending the result back to the blockchain through its own blockchain adapter. The exclusion criteria in the example in Figure 3 also has current tobacco use which is not recorded in patient A’s information. The master smart contract will notify patient A that there is a potential clinical trial that he/she might be eligible to participate in and need his/her authentication for the sponsor to access his/her EHR to double check with the details. The sponsor can communicate with the clinical sites which patient A visited before to check whether he/she uses tobacco currently with patient A’s consent by E-signature using the private key. The clinical site will perform precise checking for the sponsor.

Figure 3. An example of matching process using smart contract and ask consent to join the clinical trial
Module 2 is about how subjects send consent to join the trial chain. If patient A has fully matched the criteria, sponsor A will send a transaction to patient A to ask for enrollment, patient A can agree to join the trial using an E-signature. Then all of patient A’s primary records will be stored into the trial-based contract. Sponsor A can select the trial sites based on the density distribution after collecting all the enrolled subjects’ geographical information. The patients still need to go to the clinical sites to sign all paperwork and proceed with the trial. The sponsor can use the trial-based contract which can only be accessed by the participants of this trial publishing the trial detail and announcement.

We have also built a web-based Graphical User Interface (GUI) on the RPC servers for users to better interact with the system instead of using plain command codes in the blockchain console. To log in to the GUI, all users need to set up a username and a password. The username and password will be stored in the local RPC and mapped to the public key and private key. This could potentially be implemented using personal biometric information to log in to the system. All the functions showed in the GUIs are sending or receiving data from the blockchain rather than a cloud database. Different roles of users will have different GUIs to operate the system. The clinical site’s GUI will have the functions as input the patient’s primary visit records to the smart contract; check ongoing clinical trial as a trial site; check requests from sponsor to check on specific patient’s eligibility and send the result back to the blockchain (Figure 4(a)). The potential matched clinical trials requesting authentication will be displayed in the patient’s GUI. After the patient clicks “approve”, a transaction signed by the patient’s private key will be sent to empower the sponsor to request precise matching from the clinical sites. Patients can also check their basic information and the visiting records but cannot be changed. Patients can also check enrolled clinical trial information the same as the trial sites (Figure 4(b)). The sponsor’s GUI will have a list of ongoing, recruiting and completed clinical trials. For the recruiting clinical trials, they will show the list of matched subjects. Once the request is sent, it will deliver a transaction to the patient through the sponsor’s account requesting authentication (Figure 4(c)). The authority will have all the clinical trial lists and the trial details (Figure 4(d)).

Figure 4. GUIs for (a) clinical sites to input primary records and receive requests from sponsors, (b) patients to receive notifications and authenticate sponsors, (c) sponsors to request a precise match for potential subjects, (d) authority to monitor all the trials

Simulation

To test the feasibility and efficiency of the system, we simulated the recruitment process which started from the moment that the authority registered the trial criteria to the master smart contract, and continuing to the end of the
process when recruited subjects were added to the trial-based contract. We have randomly picked 10 currently recruiting clinical trials (NCT03778931, NCT03252431, NCT03367572, NCT02068092, NCT03152929, NCT03354611, NCT01185132, NCT03523585, NCT03529110, NCT03200704). Criteria such as specific medication use vary from different clinical trials. It is doable to code all the criteria into the master smart contract. To demonstrate the feasibility of using the system for recruitment, we only select the frequent criteria among those 10 trials for the simulation as listed in Table 3. All the selected frequent criteria can be matched from the auto-matching function directly. Five blockchain adapters are set up using the Intel NUC machines: (1) Authority (simulated authority) node as the starting node, (2) sponsor node, and (3) three different clinical sites. We have set up 2,000 synthesized patient accounts on each clinical site’s node. We use the data from Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute for the simulated cases. The SEER database contains cancer incidence and survival data covering 34.6% of the U.S. population. For our simulation, we randomly picked 6,000 patients with breast cancer from the SEER database and evenly distributed them into three clinical sites.

Table 3. Frequent criteria for breast cancer clinical trials among ten selected recruiting trials with frequency

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>1. 18 years and older (9 times)</th>
<th>2. Female (6 times)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3. Diagnosed as breast cancer (7 times)</td>
<td>4. Negative metastatic involvement (3 times)</td>
</tr>
<tr>
<td>Exclusion criteria</td>
<td>1. Stage IV cancer (8 times)</td>
<td>2. Pregnant or breastfeeding (8 times)</td>
</tr>
<tr>
<td></td>
<td>3. Persistent malignant (4 times)</td>
<td>4. Breast implants (3 times)</td>
</tr>
</tbody>
</table>

We have created one database on each clinical site’s adapter. Before the simulation process, we have written scripts to populate all the patients’ previous primary diagnoses and treatments into the master smart contract through each clinical site’s administrator’s account. We have manually input the inclusion/exclusion criteria to the master smart contract sending the consensus transaction from their patient account. We created a control script in each adapter to only send five transactions from different patient IDs to the blockchain every second. All the patients have been successfully registered in the trial-based contract without breaking the chain. Figure 5 shows the results of calling the master smart contract and trial-based contract’s functions through the trial sponsor’s account as: (1) checking the total matched patients for the trial NCT103200704; (2) getting the trial-based contract address (the ABIs of trial-based contracts are the same and pre-stored in the adapters for deployment); (3) checking whether the address of the trial-based contract is matched with the address stored in the master contract; (4) receiving the amount of enrollment for this trial; (5) obtaining the geographical information of all enrolled subjects (only showing part of the 100 enrolled subject’s zip codes) for the trial NCT103200704. The results are from the blockchain console that shows the response of calling smart contracts functions from plain codes. Users

Results

Due to the limited information of the SEER database, we cannot check the pregnancy in the exclusion criteria or whether the breast cancer is persistent malignant. After executing the auto-matching function in the master smart contract, 1,145 patients out of 6,000 patients are matched in 2.13 seconds. We have used SQL to query the databases on the same criteria and got the same results. We have randomly selected 100 matched patients to join the trial smart contract using the trial address stored in the master smart contract by sending the consensus transaction from their patient account. We created a control script in each adapter to only send five transactions from different patient IDs to the blockchain every second. All the patients have been successfully registered in the trial-based contract without breaking the chain. Figure 5 shows the results of calling the master smart contract and trial-based contract’s functions through the trial sponsor’s account as: (1) checking the total matched patients for the trial NCT103200704; (2) getting the trial-based contract address (the ABIs of trial-based contracts are the same and pre-stored in the adapters for deployment); (3) checking whether the address of the trial-based contract is matched with the address stored in the master contract; (4) receiving the amount of enrollment for this trial; (5) obtaining the geographical information of all enrolled subjects (only showing part of the 100 enrolled subject’s zip codes) for the trial NCT103200704. The results are from the blockchain console that shows the response of calling smart contracts functions from plain codes. Users
will use the GUIs (Figure 4(b)) rather than the blockchain console to send requests and receive results in practical use. The NCT103200704_contract which is a trial-based contract shown in Figure 5 is automatically generated by the master smart contract.

Discussion

The blockchain features are a good fit for the clinical trial recruitment process. From the simulation we have done, all the users can see all the recruiting trials’ information. Sponsors and clinical sites need to get the authority’s approval to join the blockchain system and all the patients’ identities are verified by the clinical sites. The transactions are public auditable and also under the surveillance of the authority. The data component in the transactions is encrypted and can only be decrypted by a certain group of users. These features ensure the authenticity of the clinical trial, data security of the transactions and the accuracy of data exchange that has occurred during the clinical trial. After integrating with the smart contract functions, the blockchain system becomes more feasible for recruitment. The auto-matching function is expected to provide the patients an efficient tool to search potential clinical trials. Only auto-matched patients will get notifications from the sponsors. The patients can comprehend the details of the clinical trial after receiving the notification. The auto-matching process also saves time for the patients to understand the complex recruitment protocol. This process shows how blockchain solves the issue of lack of awareness of the eligible clinical trials. Since the criteria is inputted by the authority and all the clinical sites share the same version of the protocol, there won’t be an insufficient trial protocol. Using the blockchain system, the patients only need to opt-in and wait for notifications of the matched potential eligible clinical trials.

The trial-based contract mechanism optimizes clinical trial management. Only the trial participants can access the data in the trial-based contracts. Each trial-based contract is isolated from each other. From the simulation results, we have narrowed down the patients’ selection pool. Selected patients have been successfully added to the trial-based contract after sending their consent using a private key. The transaction processing time depends on the block generating rate which is defined in the genesis block file. In our system, a new block will be generated at around every 30 seconds. The time consumption is acceptable for the subject matching process, solving the issue for clinical sites matching with potential subjects.

The sponsor has received the subjects’ list after all the subjects have been registered to the trial-based contract. Then the sponsor can get the subjects’ geographical distribution by simply calling the smart contract function. Then the sponsor can select trial sites after considering the subjects’ geographical distribution. This approach could potentially reduce some opportunities for clinical sites to join the clinical trials but provides benefits for patients to access the clinical trial easier and sponsors to set efficient sites for the recruitment.

Our design also provides potential solutions to the current challenges for healthcare applications involving the blockchain. The username and password setting is a potential solution to the key management vulnerability. The user can also contact the authority user to rebuild a new account and remove the original one. To empirically prove that only legitimate trials can be accessed and matched with qualified patients, we have intentionally tested malicious behaviors, such as manually changing patient’s record to meet a trial criterion, registering a fake clinical trial that does not exist in the authority’s database through a sponsor’s account, and executing auto-matching process through a clinical site’s node, we found that all of these transactions were rejected by the smart contract as expected. These experiments ensure that only transactions follow the rules listed in smart contracts will be executed.

The user can audit all the transactions theoretically through the blockchain console. However, there are three reasons that there will not be any loss of privacy: (1) the data contained in the transactions sent to the smart contract cannot be decrypted by users; (2) users can only see that all the transactions are sent to different blockchain addresses but cannot know the receivers’ identities or decrypt the data contained in the transactions; and (3) users can only call the smart function to retrieve their own data stored in the smart contract; other actions will be discarded automatically.

To demonstrate the scalability of our implementation, we have simulated a clinical trial recruitment process which contains one request from a sponsor with a 2.13 second transaction time and 100 consensus transactions, as discussed previously in the Results section, from 100 patient accounts with a 24.69 second total transaction time. Using the latest report of global participation in clinical trials by the FDA, there were 131,749 total participants during 2015-2016. Our simulation result with a controlled input transaction frequency for five transactions per second through the RPC node leads us to project that it would take several hours for both matching requests and sending the consensus for the entire year. It is noteworthy to mention that our system setting could avoid Ethereum’s scalability limitation by spacing the transactions, but it is not to solve the known scalability issue of Ethereum.
Conclusion and Future work

Through our simulation process, our blockchain model shows features that can tackle the clinical trial recruitment issues. Using the master smart contract to match patients and trial-based contracts to manage the clinical trials can optimize the recruitment process as timesaving, identifying all potential subjects, patient empowerment, and the authority’s surveillance. Trial-based contracts can be used for EHR collection for subjects during the clinical trial. Blockchain features can ensure the data provenance is clinical sites, data consistency over time, data security that can only be decrypted by certain users and patient privacy. Our future work is to further evaluate Ethereum’s validating mechanism which could fundamentally solve the scalability issue. Our extended study will be to add an artificial intelligence component to our system for the real-time detection of significant adverse events of the collected EHR during a clinical trial.

Acknowledgment

This research is supported in part by the University of Missouri Informatics and Data Science Research Initiatives (YZ and CRS), MOST108-2634-F-468-001 from the Ministry of Science and Technology through Pervasive Artificial Intelligence Research (PAIR) Labs Taiwan, and MOST 106-2632-E-468-002 (ZS, YWC, and JT).

References


28. Institute NC. Overview of the SEER Program.
Introduction—The use of clinical grouping software provides substantial value to end users by providing the capacity to facilitate dimensional reduction by transforming voluminous sets of granular International Classification of Disease (ICD) codes and grouping them into clinical groups at higher levels of representation. Many of the clinical data grouping software tools are commercially produced with substantial costs associated with licensing and use, however, several tools have available academic or non-profit organizational discounts. One non-commercial alternative for clinical data grouping is the Clinical Classifications Software (CCS) which is freely available and supported by the Healthcare Cost and Utilization Project (HCUP) which is a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ).

Prior published work on the CCS has included use in a number of studies which use the software to deal with the highly granular data in clinical encounter data including over 14,000 codes in ICD-9-CM with 3900 procedure codes. In ICD-10-CM/PCS terminology there are over 69,800 diagnosis codes and 71,900 procedural codes(1). Such large and granular datasets provide a great deal of information with high levels of detail, but their breadth makes their use in modeling difficult due to their high dimensionality. The CCS provides mappings of the ICD codes to a set of 285 mutually exclusive categories using the single level coding for diagnoses and 231 mutually exclusive categories for procedures. The single level coding are of primary interest since they can facilitate risk adjustment and diagnostic ranking work for research or software applications(1). The HCUP website for the beta version of the CCS tool used for ICD10 indicates there are no publications on the CCS beta version making this an important area of research work.

Methods – The CCS was analyzed using 2 years of National Inpatient Sample data using the first three quarters of 2015 and the first 3 quarters of 2016 to provide seasonally matching data sets that included ICD9-CM clinical data for the 2015 data set and ICD10-CM/PCS for the 2016 data set. Three quarters were used since the conversion to ICD10 occurred for the last quarter of 2015 (starting October 1, 2015). The 2015 data provided the baseline incidence rates for each of the CCS categories and a cutoff of 20% change from 2015 to 2016 was considered significant in the evaluation.

Results: There were a total of 54.7 million CCS events in the 2015 ICD9 data. There were a total of 55.9 million CCS events in the 2016 ICD10 data representing a 2.2% increase in the number of events. The total counts for each CCS category were also totaled and the significant results are noted in table 1. A total of 62 CCS categories had change from baseline of 20% or greater in the number of events among the 285 different diagnostic categories. A total of 15 categories had a change of greater than 50% from baseline with 9 of the 15 with larger numbers of CCS events after ICD10 implementation.

Discussion – The change in clinical coding from ICD9 to ICD10 has created a number of issues with previously established clinical software and research methods which has been in use in the United States. Conversion involves identifying clinically equivalent codes which exist in both terminology systems, however, the larger number of codes with greater granularity provide greater capacity with ICD10. This increased knowledge representation may not provide equivalent representation before and after the implementation occurred. Additional work may be needed to adjust the coding groups in the beta version of CCS to insure correct knowledge representation is the CCS software with a particular focus on those categories with the greatest changes in events before and after ICD10 implementation.
<table>
<thead>
<tr>
<th>CCS_CATEGORY</th>
<th>CCS_CATEGORY_DESCRIPTION</th>
<th>2015 Count</th>
<th>2016 Count</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>258</td>
<td>Other screening for suspected conditions (not mental disorders or infectious disease)</td>
<td>5623</td>
<td>116243</td>
<td>1967.3%</td>
</tr>
<tr>
<td>661</td>
<td>Substance-related disorders</td>
<td>449628</td>
<td>1154106</td>
<td>156.7%</td>
</tr>
<tr>
<td>57</td>
<td>Immunity disorders</td>
<td>13835</td>
<td>27363</td>
<td>97.8%</td>
</tr>
<tr>
<td>656</td>
<td>Impulse control disorders NEC</td>
<td>6685</td>
<td>13184</td>
<td>97.2%</td>
</tr>
<tr>
<td>209</td>
<td>Other acquired deformities</td>
<td>48536</td>
<td>88414</td>
<td>82.2%</td>
</tr>
<tr>
<td>125</td>
<td>Acute bronchitis</td>
<td>47862</td>
<td>80882</td>
<td>68.9%</td>
</tr>
<tr>
<td>50</td>
<td>Diabetes mellitus with complications</td>
<td>536361</td>
<td>855594</td>
<td>59.5%</td>
</tr>
<tr>
<td>228</td>
<td>Skull and face fractures</td>
<td>27383</td>
<td>42500</td>
<td>55.2%</td>
</tr>
<tr>
<td>195</td>
<td>Other complications of birth; puerperium affecting management of mother</td>
<td>419582</td>
<td>633468</td>
<td>51.0%</td>
</tr>
<tr>
<td>133</td>
<td>Other lower respiratory disease</td>
<td>287884</td>
<td>409363</td>
<td>42.2%</td>
</tr>
<tr>
<td>655</td>
<td>Disorders usually diagnosed in infancy childhood or adolescence</td>
<td>12647</td>
<td>17078</td>
<td>35.0%</td>
</tr>
<tr>
<td>231</td>
<td>Other fractures</td>
<td>111073</td>
<td>146712</td>
<td>32.1%</td>
</tr>
<tr>
<td>119</td>
<td>Varicose veins of lower extremity</td>
<td>8298</td>
<td>10955</td>
<td>32.0%</td>
</tr>
<tr>
<td>243</td>
<td>Poisoning by nonmedicinal substances</td>
<td>6452</td>
<td>8492</td>
<td>31.6%</td>
</tr>
<tr>
<td>252</td>
<td>Malaise and fatigue</td>
<td>103723</td>
<td>136109</td>
<td>31.2%</td>
</tr>
<tr>
<td>230</td>
<td>Fracture of lower limb</td>
<td>61900</td>
<td>80955</td>
<td>30.8%</td>
</tr>
<tr>
<td>64</td>
<td>Other hematologic conditions</td>
<td>24791</td>
<td>32145</td>
<td>29.7%</td>
</tr>
<tr>
<td>177</td>
<td>Spontaneous abortion</td>
<td>1484</td>
<td>1923</td>
<td>29.6%</td>
</tr>
<tr>
<td>41</td>
<td>Cancer; other and unspecified primary</td>
<td>14983</td>
<td>19348</td>
<td>29.1%</td>
</tr>
<tr>
<td>236</td>
<td>Open wounds of extremities</td>
<td>38634</td>
<td>49455</td>
<td>28.0%</td>
</tr>
<tr>
<td>224</td>
<td>Other perinatal conditions</td>
<td>424775</td>
<td>543136</td>
<td>27.9%</td>
</tr>
<tr>
<td>216</td>
<td>Nervous system congenital anomalies</td>
<td>14976</td>
<td>19104</td>
<td>27.6%</td>
</tr>
<tr>
<td>259</td>
<td>Residual codes; unclassified</td>
<td>2569369</td>
<td>3188124</td>
<td>24.1%</td>
</tr>
<tr>
<td>203</td>
<td>Osteoarthritis</td>
<td>461921</td>
<td>567876</td>
<td>22.9%</td>
</tr>
<tr>
<td>650</td>
<td>Adjustment disorders</td>
<td>26175</td>
<td>32109</td>
<td>22.7%</td>
</tr>
<tr>
<td>240</td>
<td>Burns</td>
<td>24362</td>
<td>29739</td>
<td>22.1%</td>
</tr>
<tr>
<td>662</td>
<td>Suicide and intentional self-inflicted injury</td>
<td>107244</td>
<td>130888</td>
<td>22.0%</td>
</tr>
<tr>
<td>197</td>
<td>Skin and subcutaneous tissue infections</td>
<td>247379</td>
<td>301765</td>
<td>22.0%</td>
</tr>
<tr>
<td>229</td>
<td>Fracture of upper limb</td>
<td>51593</td>
<td>62749</td>
<td>21.6%</td>
</tr>
<tr>
<td>127</td>
<td>Chronic obstructive pulmonary disease and bronchiectasis</td>
<td>658436</td>
<td>791865</td>
<td>20.3%</td>
</tr>
<tr>
<td>82</td>
<td>Paralysis</td>
<td>102104</td>
<td>122669</td>
<td>20.1%</td>
</tr>
<tr>
<td>10</td>
<td>Immunizations and screening for infectious disease</td>
<td>690792</td>
<td>551401</td>
<td>-20.2%</td>
</tr>
<tr>
<td>227</td>
<td>Spinal cord injury</td>
<td>10800</td>
<td>8605</td>
<td>-20.3%</td>
</tr>
<tr>
<td>219</td>
<td>Short gestation; low birth weight; and fetal growth retardation</td>
<td>130577</td>
<td>103674</td>
<td>-20.6%</td>
</tr>
<tr>
<td>123</td>
<td>Influenza</td>
<td>32999</td>
<td>25508</td>
<td>-22.7%</td>
</tr>
<tr>
<td>181</td>
<td>Other complications of pregnancy</td>
<td>492992</td>
<td>373181</td>
<td>-24.3%</td>
</tr>
<tr>
<td>104</td>
<td>Other and ill-defined heart disease</td>
<td>59303</td>
<td>48813</td>
<td>-24.4%</td>
</tr>
<tr>
<td>151</td>
<td>Other liver diseases</td>
<td>412477</td>
<td>311319</td>
<td>-24.5%</td>
</tr>
<tr>
<td>199</td>
<td>Chronic ulcer of skin</td>
<td>306206</td>
<td>230999</td>
<td>-24.6%</td>
</tr>
<tr>
<td>28</td>
<td>Cancer of other female genital organs</td>
<td>4582</td>
<td>3452</td>
<td>-27.9%</td>
</tr>
<tr>
<td>171</td>
<td>Menstrual disorders</td>
<td>31075</td>
<td>22861</td>
<td>-26.4%</td>
</tr>
<tr>
<td>148</td>
<td>Peritonitis and intestinal abscess</td>
<td>38951</td>
<td>28140</td>
<td>-27.8%</td>
</tr>
<tr>
<td>248</td>
<td>Gangrene</td>
<td>23824</td>
<td>16992</td>
<td>-28.7%</td>
</tr>
<tr>
<td>654</td>
<td>Developmental disorders</td>
<td>55084</td>
<td>38636</td>
<td>-29.9%</td>
</tr>
<tr>
<td>94</td>
<td>Other ear and sense organ disorders</td>
<td>129535</td>
<td>89092</td>
<td>-31.2%</td>
</tr>
<tr>
<td>111</td>
<td>Other and ill-defined cerebrovascular disease</td>
<td>36964</td>
<td>24905</td>
<td>-32.6%</td>
</tr>
<tr>
<td>178</td>
<td>Induced abortion</td>
<td>572</td>
<td>374</td>
<td>-34.6%</td>
</tr>
<tr>
<td>43</td>
<td>Malignant neoplasm without specification of site</td>
<td>16010</td>
<td>10349</td>
<td>-35.4%</td>
</tr>
<tr>
<td>108</td>
<td>Congestive heart failure; nonhypertensive</td>
<td>1131454</td>
<td>720432</td>
<td>-36.3%</td>
</tr>
<tr>
<td>242</td>
<td>Poisoning by other medications and drugs</td>
<td>41549</td>
<td>26420</td>
<td>-36.4%</td>
</tr>
<tr>
<td>244</td>
<td>Other injuries and conditions due to external causes</td>
<td>322406</td>
<td>198173</td>
<td>-38.5%</td>
</tr>
<tr>
<td>2</td>
<td>Septicemia (except in labor)</td>
<td>703502</td>
<td>430202</td>
<td>-38.8%</td>
</tr>
<tr>
<td>179</td>
<td>Postabortion complications</td>
<td>956</td>
<td>553</td>
<td>-42.2%</td>
</tr>
<tr>
<td>663</td>
<td>Screening and history of mental health and substance abuse codes</td>
<td>3460553</td>
<td>804550</td>
<td>-44.9%</td>
</tr>
<tr>
<td>87</td>
<td>Retinal detachments;ectas;vascular occlusion;and retinopathy</td>
<td>76653</td>
<td>38931</td>
<td>-49.2%</td>
</tr>
<tr>
<td>204</td>
<td>Other non-traumatic joint disorders</td>
<td>225524</td>
<td>114229</td>
<td>-49.3%</td>
</tr>
<tr>
<td>670</td>
<td>Miscellaneous mental health disorders</td>
<td>67618</td>
<td>33402</td>
<td>-50.6%</td>
</tr>
<tr>
<td>256</td>
<td>Medical examination/evaluation</td>
<td>104799</td>
<td>50775</td>
<td>-51.6%</td>
</tr>
<tr>
<td>241</td>
<td>Poisoning by psychotropic agents</td>
<td>25577</td>
<td>9915</td>
<td>-61.2%</td>
</tr>
<tr>
<td>220</td>
<td>Intraventricular hypoxia and birth asphyxia</td>
<td>4399</td>
<td>1331</td>
<td>-69.7%</td>
</tr>
<tr>
<td>156</td>
<td>Nephritis; nephrosis; renal sclerosis</td>
<td>66188</td>
<td>17460</td>
<td>-73.6%</td>
</tr>
<tr>
<td>254</td>
<td>Rehabilitation care; fitting of prostheses; and adjustment of devices</td>
<td>64038</td>
<td>673</td>
<td>-98.9%</td>
</tr>
</tbody>
</table>

The Impact of Transitioning from Availability of Outside Records within EHR to Integration of Local and Outside Records within EHR

Julia Adler-Milstein, PhD¹ and Michael Wang, MD¹
¹University of California San Francisco, San Francisco, CA.

Introduction
Substantial effort over the past decade has resulted in greater electronic availability of outside records for frontline clinicians. However, most approaches to enabling such availability require clinicians to go outside of their local EHR (e.g., by logging in to a community longitudinal record) or to go to a separate tab in their local EHR that houses outside records. Under either approach, clinician workflow is interrupted and cognitive effort is required to marry local EHR data with data available in outside records. Newer approaches seek to comingle local EHR data with data from outside records, such that clinicians stay within their workflow and are presented with an integrated list of encounters, lab results, problems, medications, etc. Given the substantial work required to achieve this “last mile” step of integrating data from local and outside records, it is critical to assess the impact of such integration. We therefore undertook the first-ever study to specifically investigate whether the frequency of outside record viewing increases after such integration as well as characterize for which types of encounters, viewers, and patients any increases accrued.

Methods
On July 11, 2018 UCSF Health, a large academic medical center, turned on a new feature that created a single, integrated list of local (from UCSF Health) and external (from any other Epic-based health system) encounters presented in the Chart Review (CR) tab. Prior to this switch, outside records were only available through the CareEverywhere (CE) tab, a section of the EHR exclusive to outside records. The new feature was supplementary, such that there was no change to user ability to view outside records via the CareEverywhere tab.

To study the impact of this change on patterns of frontline clinician viewing of outside records, we studied the 12-month period (calendar year 2018) centered on the new feature go-live date (July 11, 2018). We first conducted an interrupted time series analysis (ITSA) to assess overall changes in level and/or trend of volume of outside record views as captured in the Clarity User Event Log table. We repeated this analysis using the granular nature of the Event Log tables to measure how often an outside records view was preceded by opening Chart Review (CR) or CareEverywhere since we hypothesized that most/all of any increases would come from the new CR pathway. Next, we performed stratified analyses by user type (Residents, Attendings, Nurses, Others) and by encounter type (Emergency, Inpatient, Encounter, Other). Finally, we compared user and patient demographics for encounters with outside record viewing in the three month pre (March-May) versus post periods (August-October) to capture any differences in who/for whom outside records were viewed (e.g., younger patients).

Results

Changes in Volume of Outside Record Viewing

Figure 1. Number of Outside Record Views Before and After Switch to Integrated Local/Outside Records

Figure 1 displays the total weekly number of outside record viewing events, with a visible increase in week 28 when the new feature was switched on. In our interrupted time series analysis, there was a statistically significant increase in level of 22,646 views (p<0.001). There was also an increasing pre-period slope (244 views per week; p<0.01) and a decreasing post-period slope (-569 views per week; p<0.05). When we performed the ITSA for views preceded by opening Chart Review (CR) or CareEverywhere since we hypothesized that most/all of any increases would come from the new CR pathway. Next, we performed stratified analyses by user type (Residents, Attendings, Nurses, Others) and by encounter type (Emergency, Inpatient, Encounter, Other). Finally, we compared user and patient demographics for encounters with outside record viewing in the three month pre (March-May) versus post periods (August-October) to capture any differences in who/for whom outside records were viewed (e.g., younger patients).
in viewing level, revealing that increases were not concentrated among specific user or encounter types. All pre-period slopes were positive and significant, but many post-period slopes were no longer significant, indicating that specific subgroups (attendings, emergency and outpatient encounters) did not have an effect decay post-intervention.

**Changes in Characteristics of Users and Patients with Outside Record Viewing**

Table 1 presents the characteristics associated with encounters with outside records views in the pre-versus-post periods. Across encounter types, there were consistent increases in the number of outside records viewed per encounter as well as the number of unique viewers. We also observed some differences in viewer type (with a trend towards greater viewing among nurses) and some general trends (though many were not statistically significant) towards lower imaging utilization that could result from greater viewing of outside records as has been found in previous studies.

**Table 1. Viewer and Patient Demographics Before and After Switch to Integrated Local/Outside Records**

<table>
<thead>
<tr>
<th>Viewer Characteristics</th>
<th>Pre (Mar-June)</th>
<th>Post (Aug-Nov)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total OR Viewed</td>
<td>0.50</td>
<td>1.00</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Avg. # OR Viewed</td>
<td>20.1</td>
<td>20.3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Viewer Characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending (%)</td>
<td>15.3</td>
<td>15.1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Resident (%)</td>
<td>41.0</td>
<td>41.0</td>
<td>0.001</td>
</tr>
<tr>
<td>Nurse (%)</td>
<td>8.9</td>
<td>9.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Other (%)</td>
<td>34.8</td>
<td>35.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Patient Characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female %</td>
<td>0.50</td>
<td>0.51</td>
<td>0.05</td>
</tr>
<tr>
<td>Average Age (Yrs)</td>
<td>46.6</td>
<td>47.7</td>
<td>0.01</td>
</tr>
<tr>
<td>Average # 2018 IP Visits</td>
<td>9.03</td>
<td>9.06</td>
<td>0.001</td>
</tr>
<tr>
<td>Average # 2018 OP Visits</td>
<td>3.95</td>
<td>4.09</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average # 2018 ED Visits</td>
<td>3.72</td>
<td>3.73</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Insurance Class</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial (%)</td>
<td>34.7</td>
<td>34.9</td>
<td>0.37</td>
</tr>
<tr>
<td>Medicaid (%)</td>
<td>24.5</td>
<td>24.4</td>
<td>0.05</td>
</tr>
<tr>
<td>Medicare (%)</td>
<td>37.4</td>
<td>37.8</td>
<td>0.001</td>
</tr>
<tr>
<td>Self/Pay (%)</td>
<td>2.6</td>
<td>2.0</td>
<td>0.001</td>
</tr>
<tr>
<td>Other (%)</td>
<td>0.0</td>
<td>0.1</td>
<td>0.001</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American (%)</td>
<td>20.0</td>
<td>19.6</td>
<td>0.10</td>
</tr>
<tr>
<td>Asian (%)</td>
<td>13.6</td>
<td>13.4</td>
<td>0.01</td>
</tr>
<tr>
<td>Caucasian (%)</td>
<td>50.6</td>
<td>51.0</td>
<td>0.05</td>
</tr>
<tr>
<td>Other/Declined (%)</td>
<td>23.2</td>
<td>23.4</td>
<td>0.01</td>
</tr>
<tr>
<td>Imaging Utilization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average # Xray Ordered</td>
<td>3.71</td>
<td>3.31</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average # CT Ordered</td>
<td>1.03</td>
<td>0.98</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average # MR Ordered</td>
<td>0.11</td>
<td>0.09</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Clinical Outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of Stay (days)</td>
<td>8.14</td>
<td>7.82</td>
<td>0.02</td>
</tr>
<tr>
<td>Average Charlson Index</td>
<td>1.78</td>
<td>1.74</td>
<td>0.42</td>
</tr>
</tbody>
</table>

**Discussion**

While limited by a single-site, single EHR-vendor setting as well as the potential for simultaneous event confounding, we conducted a novel study that found that a user interface change that presented integrated local and outside records in the Chart Review tab resulted in a large increase in the level of outside record viewing. Increases were spread across user and encounter types, indicating the widespread impact of this change on ensuring that frontline clinicians have ready access to outside records. However, we did observe some reduction in post-period trend, suggesting a slow effect decay over time. Given the substantial federal investment in interoperability, our results indicate that the benefits will not be maximized unless we specifically focus attention on last mile issues of integrating external data, rather than simply ensuring it is electronically available. This is a blindspot in current interoperability policies, including Promoting Interoperability Program criteria and 21st Century Cures. It is therefore incumbent upon health systems and EHR vendors to push for integration of outside data into clinician workflows in order to mobilize data to improve patient care. Future work tying integration and resulting increased viewing to outcomes will help bolster such efforts.

**References**

An Evaluation of an Artificial Intelligence Clinical Trial Matching System in Australian Lung Cancer Patients

Marliese Alexander, B.Pharm, PhD\(^1\), Benjamin Solomon, MBBS, PhD\(^1\), David L Ball, MBBS, MD\(^1\), Mimi Sheerin,\(^2\) Irene Dankwa-Mullan MD MPH,\(^2\) Gretchen Purcell Jackson, MD PhD,\(^2\) Dishan M Herath, MBBS\(^3\),
\(^1\)Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia; \(^2\)IBM Watson Health, Cambridge MA, USA

**Introduction:** Prospective clinical trials are the gold standard for assessing the potential harms and benefits of new cancer treatments. However, clinical trial recruitment is challenging and time-consuming.\(^1\) As of 2015, only about 6% of patients with a cancer diagnosis in the state of Victoria in Australia were recruited to clinical trials, with rates unchanged over more than a decade. A key deterrent is the tedious manual process of matching patients to clinical trials. IBM\(^\circledR\) Watson for Clinical Trial Matching (CTM) is a software platform developed to identify potential trials for an individual patient or potential trial candidates for an individual trial. CTM uses natural language processing (NLP) to intake trial and patient information from unstructured sources and match patients to trials for which they are eligible. The objective of this retrospective study was to evaluate the performance of CTM for lung cancer patients in an Australian cancer hospital.

**Methods:** Information about 10 phase I–III cancer clinical trials registered on clinicaltrials.gov and open to lung cancer patients at Peter MacCallum Cancer Centre, a specialist cancer hospital in Melbourne, Australia, was processed by CTM. Lung cancer patients were selected in reverse chronological order from an IRB approved research database, with data current for all patients as of August 2018. CTM-processed eligibility criteria were checked and refined by two clinicians (medical oncologist and pharmacist) who entered de-identified patient attributes such as histological diagnosis, stage and prior therapies into CTM. A timed query was executed for each patient to assess eligibility for potential trials, with trials classified as “Exclude” (patient not eligible) or “Consider” (patient potentially eligible). A gold standard for trial eligibility was determined for each patient and the 10 cancer trials by independent review of patient attributes entered into CTM (not the full medical record) by two clinicians with discrepancies discussed to achieve consensus. Accuracy, recall and precision of CTM trial classification was measured.

**Results:** A total of 102 lung cancer patients were included in the study. CTM evaluated a total of 7,252 patient attributes (per patient median 74, range 53-100) against 11,467 individual trial eligibility criteria (per trial median 597, range 243-4132). The median time taken for CTM to run a query and return trial results was 15.5 seconds (range 7.2-37.8 seconds). In establishing the gold standard comparator, clinician interrater agreement was high (Kappa=0.70–1.00), with disagreement due to selection error or overlooked features/criteria rather than material disagreement.

On a per-patient basis, the accuracy of CTM for eligibility classification across trials was 91.6%, recall (sensitivity) 83.3%, precision (positive predictive value) 76.9%, negative predictive value 95.7% and specificity 93.8%. When considering only trials classified as “Exclude” by CTM, accuracy was 95.7% with only 34 of 799 trials incorrectly excluded. Conversely, for trials classified as “Consider” by CTM, accuracy was 76.6%, with 52 of 221 labelled “Consider” that should have been excluded. CTM accuracy for individual trials ranged from 77%-100%, Table 1.

<table>
<thead>
<tr>
<th>Trial ID</th>
<th>Number of criteria assessed</th>
<th>% patients listed for trial consideration</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>270</td>
<td>1% CTM 2% Gold Standard</td>
<td>99%</td>
</tr>
<tr>
<td>2</td>
<td>560</td>
<td>1% CTM 1% Gold Standard</td>
<td>92%</td>
</tr>
<tr>
<td>3</td>
<td>243</td>
<td>0% CTM 0% Gold Standard</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>1010</td>
<td>23% CTM 18% Gold Standard</td>
<td>93%</td>
</tr>
<tr>
<td>5</td>
<td>313</td>
<td>3% CTM 6% Gold Standard</td>
<td>97%</td>
</tr>
<tr>
<td>6</td>
<td>4132</td>
<td>59% CTM 40% Gold Standard</td>
<td>77%</td>
</tr>
</tbody>
</table>
Among all trials, 1490 trial eligibility criteria (inclusion/exclusion) were listed as ‘not met’ by CTM (90% agreement with gold standard), 1231 were ‘met’ (96% agreement), 8088 were identified as requiring further action to make a decision and listed as ‘action needed’ (89% agreement), 136 were identified as ‘unmet modifiable’ (90% agreement), and 522 consent criteria were reviewed (81% agreement). The number of data elements and criteria varied by trial, with trial 6 representing an umbrella multi-cohort design with notably higher number of criteria, Table 1.

Discussion

Previous studies have shown that an automated clinical trials eligibility tool and CTM can reduce the time taken to screen patients for clinical trials and increase trial enrolment. At Mayo Clinic, accuracy was reported as 87.6% for 4 breast and 74.9% for 3 lung cancer protocols. Other tools show similar results but accuracy data of this type is limited. This study expands the knowledge about CTM performance to a larger number of patients and trials, and to a setting outside of the United States. In our unselected patient cohort, CTM software was able to reliably exclude ineligible patients from trial consideration (>95% accuracy) but was less accurate for determining who was truly eligible (77%). One contributing factor was that for the 102-patient cohort, 8088 data items were identified as requiring further action (data input or clinician interpretation). This study has several limitations. First, although CTM is capable of processing structured and unstructured information from an electronic health record (EHR), only the matching components of the CTM system were evaluated because the EHR was not integrated with CTM for this study. Second, the study included a relatively small number of patients at a single center. Strengths of the study are a rigorous gold standard for eligibility with consensus agreement of two clinicians who had a high interrater reliability.

Conclusion

This study demonstrated that CTM allows efficient and reliable screening of Australian lung cancer patients for clinical trials, with excellent accuracy in exclusion and good performance in assessing potential eligibility. However, clinician input and oversight are still required for assessing nuances of patient characteristics against individual criteria. CTM demonstrates promise as a tool to pre-screen a large patient cohort to identify subjects suitable for further assessment.

References

Learning to Estimate Nutrition Facts from Food Descriptions

Hadi Amiri, PhD; Andrew L. Beam, PhD; Isaac S. Kohane, MD, PhD
Department of Biomedical Informatics
Harvard Medical School, Boston, MA 02115

Introduction.

Risk factors such as body weight, blood pressure, and blood cholesterol can help people make informed decisions about their health promotion efforts. Food choices are among the most effective of these efforts which can help preventing chronic diseases, such as heart disease, diabetes, stroke, and certain cancers[6]. Since different foods provide different energy and nutrients, healthy eating requires monitoring the nutrients that we consume. In addition, food offers new perspectives on topical challenges in Natural Language Processing and Computer Vision centering around computational models to extract nutrition facts from food-relevant textual content[1] or find representations that are robust to occlusion and deformation in processing of food images[2]. In fact, there is an emerging literature investigating Food Computing which aims to acquire and analyze food data from disparate sources for recommending and monitoring food consumption as well as addressing food-related issues in medicine, biology, gastronomy, and agronomy[7]. The availability of large-scale food datasets and the above recent advances in Food Computing can transform the way that individuals consume food. Established in the literature is the ability to match foods with databases that contain nutrition facts[1]. These approaches are effective for foods that exists in databases, but lack the ability to deal with large amount of new foods that don’t exist in such databases; as reported in[5], the average number of new foods per year is slightly less than 20K. In addition, there is a dearth of evidence as to whether learning food ingredients can help more accurate estimation of nutrition facts. Therefore, the aim of our study is two fold: (1) to develop effective computational models that accurately estimate nutrition facts of any given food, and (2) to investigate if computational modeling of food ingredients can help better estimation of nutrition facts.

Data Description. The USDA branded food products database[4] contains food description, nutrition facts, and ingredients for a large number of foods which are voluntarily supplied by food industry organizations to USDA. USDA standardizes the reported nutrition facts by calculating nutrient values per 100 grams from those values provided per serving. The dataset contains about 237K food items, 40 nutrition fact types, and 100K ingredient types respectively. It exhibits a power law distribution, $c \times \exp(-0.14x)$, as some nutrition facts match with only a small number of foods. Although the majority of foods in USDA dataset contain important nutrition facts, some food organizations provide no or partial information about their products. In addition, the dataset is updated on a yearly basis[4] and therefore many new foods do not exist in the dataset. These challenges inspires our work to develop computational models to automatically estimate nutrition facts from food descriptions.

Method. We developed multiple regression approaches including least squares Linear Regression (Linear) with L2 regularization (Ridge). In addition, the state-of-the-art approach for learning nutrition facts of foods given their descriptions was reported in[1]–a Convolutional Neural Network (CNN) which used word n-grams to match food items with USDA dataset to derive nutrition facts. We extended this approach through joint learning of nutrition facts and ingredients of given foods. In particular, as Figure 1 depicts, we developed a multi-task learning framework to enable joint learning of ingredients and nutrition facts given food descriptions. Note that nutrition facts were normalized scalars and learned separately (see Discussion), and ingredients of each food were represented by a vector of 0/1s with 1 indicating existence of a specific ingredient in the food. The shared layers were used to exploit commonalities.
and differences across tasks for more accurate learning. The network was trained by minimizing the following Mean Squared Error loss functions:

\[ L(I) = L_{\text{nutrition}}(I) + \alpha \times \beta \times L_{\text{ingredient}}(I) \]

where \( \alpha \in [0, 1] \) controlled the extent to which ingredients contributed in overall learning of the task; \( \alpha = 0 \) indicates no contribution and was considered as a baseline here (basic CNN), and the parameter \( \beta \) was used to establish a common scale for loss magnitudes across tasks\(^3\). Although, \( \beta \) could be tuned through grid search, we set \( \beta = L_{\text{nutrition}}^{0}/L_{\text{ingredient}}^{0} \) where \( L_{\text{ingredient}}^{0} \) indicates loss at first iteration. Both loss functions optimized minimum squared error (MSE). Our framework enables learning semantic relations between food items and ingredients, e.g. learning that “roasted” foods should have “oil” as their ingredients, as well as semantic relations between foods and nutrition facts, e.g. learning that “rice” generally has high calories. Such relations are important indicators for accurate prediction of nutrition facts as they capture aspects of nutrients which may not be effectively represented in food descriptions.

**Results.** Our multi-task learning framework was trained and tested on each nutrition fact separately. For each nutrition fact, we partitioned food items into training data (80%), development data (10%) for parameter tuning, and test data (10%) for evaluation. We used grid search to optimize \( \alpha \) for each nutrition fact using development data, then the resulting best model for each \( \alpha \) was applied to the test data. Models were compared based on Coefficient of Determination \( (R^2 \text{ score} \in (-\infty, 1], \text{where } R^2 = 1 \text{ indicates perfect regression}) \). Our multi-task learning framework outperformed Linear, Ridge, and basic CNN regressors on 70%, 55%, and 35% of nutrition fact categories respectively. In addition, Figure 2 shows the average \( R^2 \) performance over all nutrition fact categories for the top two best performing models, basic CNN and multi-task CNN respectively, across \( \alpha \) values. As the results show, \( \alpha = 0 \) led to \( R^2 \) of 22.40, while there existed other \( \alpha \) values, i.e. \( \alpha \in \{.1, .4, .8, .9, 1.0\} \), that further improved the performance. We attribute this improvement to our model’s ability in utilizing semantic relations between food items, and their ingredients and nutrition facts.

**Discussion.** In this work, we developed an effective regressor to accurately estimate nutrition facts of foods. Our work highlighted the importance of learning ingredients for accurate estimation of nutrition facts. Our research has high value for developing diet monitoring applications, which may generate results with significant public health impact. Future investigations might explore associations among food quantity and type with ingredients and nutrition facts. In addition, ingredients often have a hierarchical form, e.g. iodized salt, himalayan salt, and crystal salt can all be mapped to the ingredient salt, which could be utilized to create a better semantic space for ingredients. In addition, our learning framework is trained on each nutrition fact separately; joint learning of these facts might create stronger regressors.

**References**

Bridging Research and Clinical Care: Real-Time ePRO Adverse Event Reporting in a Phase 2 Breast Cancer Study

Adam Asare, PhD1,2, Smita Asare2, Amrita Basu, PhD1, Sue Dubman, MS2, Mitra Rocca, PhD3, Aheli Chattopadhyay1, Jessie Hong2, Lisa Weiss, MBA2, Ruby Singhrao, MS, CCRP1, Amy Wilson2, Garry Peterson1, Devon McCabe2, Ronak Ahir1, Jeff Matthews, PhD2, Cal Collins5, Karyn DiGiorgio, RN2, Bev Parker, MSW, PhD4, Susie Brain4, Diane Heditsuian4, Dave Mandelkern, MS2, Laura Esserman, MD, MBA1 and Michelle Melisko, MD1

1University of California, San Francisco, CA, 2Quantum Leap Healthcare Collaborative, San Francisco, CA, 3U.S. Food & Drug Administration, Silver Spring, MD, 4I-SPY Advocacy Group, 5OpenClinica, Waltham, MA

Introduction: Since the 21st Century Cures Act was signed into law in 2016, there has been a growing interest in the use of Real World Data (RWD) and Real World Evidence (RWE) in healthcare. RWE has multiple uses – it can inform clinical practice guidelines and coverage decisions, provide information on drug safety for regulators and serve as endpoints in pragmatic clinical trials. One source of RWE, patient-reported outcomes, appears particularly well-suited to supplement adverse event reporting in the clinical trial environment, as they provide a patient-centric view of severity and impact of treatment side effects on quality of life (QOL). Internet and smart phone technologies now provide the opportunity to collect patient-reported outcomes in a more timely and efficient manner.

Here we describe the development and deployment of systems and related workflows for monitoring of patient reported adverse events (AEs) and QOL within an ongoing phase II breast cancer clinical trial using electronic patient-reported outcomes (ePROs) instruments. The instruments are administered within the Investigation of Serial studies to Predict Your Therapeutic Response with Imaging And mOLecular analysis (I-SPY 2 TRIAL), a phase 2, standing, randomized controlled, multicenter trial aimed to rapidly screen and identify promising neoadjuvant chemotherapy +/- new investigational treatments in specific subgroups of women with newly-diagnosed, high-risk for recurrence, locally-advanced breast cancer (stage II/III) that has been accruing since 2010. We will provide preliminary assessments of the feasibility and impact on the patient perceived quality of care and satisfaction.

Methods: The I-SPY 2 TRIAL currently has patients enrolled and randomized across 16 clinical sites in the United States. Paper-based QOL surveys include the EORTC QLQ-C30, EORTC QLQ-BR23, and PROMIS® measures and have been distributed by paper to patients starting January 2012. In previous work within the I-SPY 2 TRIAL, PROMIS® questions included in the paper-based I-SPY 2 QOL assessment were found to be the most appropriate and most sensitive for the early breast cancer setting. A decision was made to move from paper to ePRO in 2019. In the ePRO release, PROMIS measures selected address the following areas: Anxiety, Depression, Fatigue, Physical function, Sexual Function, and Sleep. For adverse events, the National Cancer Institute (NCI) had developed a patient-reported outcomes version of the Common Terminology Criteria for Adverse Events (CTCAE™), termed PRO-CTCAE™, which consists of a series of questions that characterize the frequency, severity and/or interference of 78 symptomatic treatment toxicities across 14 different domains. The PRO-CTCAE are focused on physical functions. These include symptomatic toxicities such as pain, fatigue, nausea, and cutaneous side effects such as rash and hand-foot syndrome. To reduce the survey burden on patients, I-SPY investigators, leadership and patient advocates identified a set of PRO-CTCAE domains/items relevant to the population to include in the ePRO survey.

Beginning in July 2019, we deployed the revised ePRO surveys to the OpenClinica (OC) Participate™ test environment and leveraged OpenClinica Insight™ for alerts and periodic reports of ePRO results. The reports will be sent to clinicians and site coordinators.

At the initial visit after the patient has consented, the site coordinator will link the patient’s email and subject ID in the OC Participate™ system. This information is encrypted in the study database. Three days prior to a specified visit requiring a survey, the patient is sent an email and/or short message service (SMS) notification requesting that they complete the assigned ePRO survey. Site coordinator will monitor data completeness. Patients who have not completed the survey prior to their visit are provided with a tablet computer with internet connectivity to complete the survey in the clinic waiting room and are supported by the site’s front desk personnel. The site coordinator will
confirm patient identification and have pre-loaded the web link to the survey to minimize patient burden. The site coordinator will pre-load the providers email information to ensure visibility in the clinic. Once the responses are captured and saved, the data results are normalized and available to investigators/providers. Thresholds will be established to send alerts/triggers to providers.

At the University of California, San Francisco (UCSF) Breast Cancer Clinic study site, we are piloting an integration with the UCSF Epic Electronic Health Record (EHR) system to receive the I-SPY 2 TRIAL individual patient ePRO results over time (Fig 1) into Epic for clinician review at the point of care. Patients at the UCSF site will be re-consented/consented to allow their PRO data to enter the UCSF EHR. An HL7 interface will be implemented to exchange data between UCSF’s Admission, Discharge, Transfer/scheduling systems and the I-SPY 2 Electronic Data Capture (EDC) system to establish the patient’s identity and match the summary ePRO report with the corresponding patient record in Epic.

Figure 1: Example plot available to providers within the UCSF Breast Cancer Clinic setting showing an I-SPY 2 patient PRO baseline result in relation to other I-SPY 2 patients in the study longitudinally through the course of the study.

Results: We will present results describing the feasibility of ePRO integration into I-SPY 2 by comparing historical paper-based survey completion rates to the ePRO completion rates over a three-month period. Patient summary ePRO reports and longitudinal plot summaries as shown in Fig 1 will be written back to the UCSF Epic system. We will compare historical paper-based QOL data to QOL collected with ePRO in the control arm to assess changes in the quality and accuracy of calculating QOL measures following ePRO use.

Discussion: Collecting patient-reported clinical trial data electronically provides the opportunity to extend the use of the data beyond the clinical trial research to: (1) provide clinicians with valuable summary information about that patient’s symptoms and quality of life at the point of care; and (2) provide real-time monitoring and alerts to clinicians for patients experiencing select adverse events or declining quality of life.

At the core of this project, transitioning from paper-based surveys to electronic surveys will improve the efficiency and accuracy of collecting patient-reported data for clinical trials. More importantly, however, will be the opportunities created by integrating this data with the EHR, transforming the uses of patient-reported clinical trial data beyond just efficiency to improve quality of care, patient safety, and health outcomes.
Characterizing the Urban Opioid Epidemic Using EHR Data

Amelia J Averitt, MPH MA MPhil; Benjamin H Slovis, MD MA; Abdul Tariq, PhD; David Vawdrey, PhD; Adler J Perotte, MD MA

1Columbia University Department of Biomedical Informatics, New York, NY, USA; 2Thomas Jefferson University Department of Emergency Medicine, Philadelphia, PA, USA; 3NewYork-Presbyterian Hospital Value Institute, New York, NY, USA

ABSTRACT The opioid epidemic is a public health emergency. Unlike typical characterizations of individuals with opioid use disorder, this research leverages electronic health records (EHR) of a large, urban, academic medical center. We confirm the increase in opioid-related overdoses through the EHR data, present demographics, and analyze the data to highlight changes in medical encounters that surround the first overdose. These insights may inform abatement of the urban opioid epidemic in a targeted way.

INTRODUCTION Characterizing the opioid epidemic (OE) from a rich and inclusive data source is a keystone for its abatement. Many strategies towards abatement of the OE include prescription drug monitoring programs (PDMPs) or prescribing limitations. These methods aim to limit over-prescription of opioids, predominately at the state-level. However, important variations exist in lower-level geographic areas and at the individual level. A more informed approach to combating the OE would address opioid cases at the individual-level with consideration of geographic context. A thorough, longitudinal characterization of this cohort may support an individual-level intervention. Characterization of the OE is often done through the analysis of claims data or a manual review of clinical documentation. However, administrative claims data may be subject to coding biases and only captures billable encounters for the insured, and manual reviews of medical records is time-consuming. The electronic health record (EHR) is a rich, longitudinal data source that captures a variety of patients with greater detail. Unlike claims data, EHR data provides not only a means to uncover overall trends in overdoses (OD), but it also supports the identification of healthcare utilization trends that are common in OD patients. We present data on all non-heroin opioid ODs in the NewYork-Presbyterian Hospital (NYPH) EHR. We examine the OE in an urban environment, as recent estimates show that urban epidemic is worsening. In addition to tracking the frequency of ODs over time, we also contrast the healthcare utilization in the period just prior to and just after first OD. By characterizing patients according to patterns in their EHR, we provide another avenue that may support our understanding of the current OE in urban areas.

METHODS The NYPH EHR contains observational clinical data for 5.37 million individual subjects from 1986-2017. Patients encounters are documented in the EHR at each outpatient, inpatient (IP), and emergency room (ER) visit. Data modalities include, but are not limited to, diagnoses, clinical measurements, medications, and procedures. Case Identification. To investigate opioid OD in the NYPH EHR, we identified all non-heroin opioid OD patients between 1/1/2006 and 12/31/2015. We mapped validated codes for non-heroin ODs from ICD-9-CM to Observational Health Data Sciences and Informatics’ common data model concept IDs. Opioid OD patients (cases) who qualified for this review must have had at least one of these diagnosis codes in a hospital visit with the ER or as an IP. Trend Analysis. To confirm an increase in the number of opioid ODs at our institution, we fit a single effect Poisson regression to model the relationship between year (2006-2015) and the rate of admissions with opioid ODs out of all ER and IP admissions, regardless of diagnosis. Demographics Analysis. We present demographic data, such as age group, and sex, along with healthcare utilization, prescriptions, medical history, and death, for the subset of opioid cases. We present the same metrics over three periods of interest, (i) The Vanilla Period, which typifies the steady-state healthcare use of patients (12 to 6 months before OD); (ii) The Pre-OD Period that just precedes the first OD (6 to 0 months before OD); and (iii) The Post-OD Period that directly follows the patient’s first OD (0 to 6 months after OD). To better contextualize the opioid case cohort, we additionally present data on controls with no history of substance abuse. Controls were randomly selected to match the distribution of age and sex of the case cohort. Self-Controlled Disproportionality Analysis. To better understand patterns leading up to and directly after the first OD, we completed two experiments; (1) Pre-OD Analysis (Pre-OD Period vs Vanilla Period) to understand changes in exposures immediately before first OD; and (2) Post-OD Analysis (Post-OD Period vs Vanilla Period) to understand changes in exposures immediately after first OD. For the two experiments, we undertook disproportionality analyses (DPAs) to look at exposure signals in (i) conditions, (ii) procedures, and (iii) medications at the ingredient
hospitals, as the rate of opioid OD for IP and ERs at NYPH coincides well with the rates reported for New York City as a whole. The results of our demographics analysis (Table 1) demonstrate that a larger proportion of OD had prescriptions for analgesic drugs in all study periods compared to controls. When examining Opioid Analgesic, 8% of the control cohort held prescriptions for this medication class, while OD cases had notably increased opioid prescriptions in all study periods (16-30%). This implies that those with opioid prescriptions are at increased risk for OD.

Results of the self-controlled DPAs highlight trends that characterize the progression of our opioid case population. The DPA of procedures may indicate that a traumatic injury, intensive care treatment, or a scheduled surgical procedure took place just prior to the first OD. Similar high-odds procedures are seen just after OD, which may indicate that further traumatic injuries occur. No medications at the ingredient level were found to be significant. The DPA of conditions highlights the close relationship between drug abuse and mental health (Figure 2). The Vanilla+Pre-OD Period is marked by high pain conditions, which may suggest that ODs stem from legitimate prescriptions. The Post-OD Period is distinguished by complications of prolonged opioid use. However, in both periods, mental health appears to be strongly associated with opioid OD.

CONCLUSION The results of this analysis indicate that the continuum of the urban OE may begin with conditions associated with pain and that OD may be tied to legitimate prescriptions. Clinicians should consider that opioid dependence may develop in medically-necessary scenarios, and lead to an OD in the short-run. Informatics-based methodologies may play a role in better understanding those individuals who suffer from opioid dependency and OD, and may lead to future research and interventions that could successfully prevent morbidity and mortality associated with the OE.

REFERENCES
Prediction of Imaging Outcomes from Electronic Health Records: Pulmonary Embolism Case-Study

Imon Banerjee, Ph.D.1,2, Miji Sofela, MS.3, Timothy Amrhein, MD.4, Daniel L. Rubin, MD.2,1, Roham Zamanian, MD.,5, Matthew P. Lungren, MD.2

1Department of Radiology, Stanford University, Stanford, CA; 2Department of Biomedical Data Science, Stanford University, Stanford, CA; 3Duke University Health System, Duke University School of Medicine, NC, USA; 4Department of Neuroradiology, Duke University School of Medicine, NC, USA; 5Department of Medicine - Med/Pulmonary and Critical Care Medicine, Stanford University, Stanford, CA;

email - imonb@stanford.edu

Introduction

Pulmonary embolism (PE) is a life-threatening clinical problem and CT imaging is the current gold standard for diagnosis. In the past years, a substantial rise in the number of CT examinations for PE evaluation has been observed with a concomitant decrease in imaging yield (as low as 2 – 3% yield). Unnecessary use of CT not only carries risks due to radiation and intravenous contrast, but also the discovery of low impact incidental imaging findings which further expose patients to unneeded procedures, tests, and risks. Clinical decision support rules based on PE risk scoring models to inform CT imaging decisions have been developed but are underutilized, in part due to evolving risk factors for PE that are not included in static scoring systems (Wells, PERC, rGeneve). The purpose of this study is to design and evaluate a machine learning modeling approach for predicting PE imaging outcomes based on patient EMR data captured before the CT exam which includes demographics, vital signs (change from baseline), diagnoses, medications, lab tests, etc. In contrast with the current clinical scoring systems that rely on a very limited set of pre-selected EMR variables, our objective is to build a machine learning solution that can compute a patient-specific risk score for PE by utilizing complex correlation between hundreds of EMR variables without relying on hand crafted feature selection; we evaluate this model on intra- and extra-institutional patient data and compare to existing PE risk scoring systems.

Methods

Data

Internal dataset (SHC) – Using the STAnford medicine Research data Repository (STARR), we randomly retrieved 4512 contrast-enhanced CT chest examinations of adult patients performed between January 1, 1998 and January 1, 2016 at Stanford hospital and clinics (SHC). Each study was manually annotated by three experienced radiologists and assigned two binary class labels (PE present/absent and PE acute/chronic). Inter-rater reliability among the three raters were highly consistent for two categories, “PE presence” and “PE Acute”, with kappa scores of 0.959 and 0.969 respectively. Because we engineered the model to identify acute PE, we dropped chronic cases to generate the final annotated internal cohorts (SHC) of 3,397 annotated PE-CT exams from 3,214 unique patients (1704 women, mean (SD) age, 60.53 (+/- 19.43)). From the EMRs, we also extracted the following phenomic data and time stamps for these patients: (1) all diagnosis codes; (2) all inpatient and outpatient medications (normalized to RxNorm); (3) all laboratory data raw values; (4) all collected vital sign data (i.e. height, weight, BMI, pulse, respiration rate, systolic blood pressure, temperature, etc); (5) all demographics (i.e. age, race, gender).

External dataset (Duke) - As external dataset, we collected contrast-enhanced CT examinations of chest from Duke University Medical Center performed between January 1, 2013 and August 31, 2017. We retrieved similar set of phenomic data of these patients with encounter time-stamp details. The data were normalized according to the standards - RxNorm, IC9 code. In order to create an external annotated dataset for validation, we randomly selected 300 CT exams from Duke and performed manual annotation with the same group of radiologists. After dropping chronic case, the models were validated on 240 unique patients seen at Duke (132 women, mean (SD) age, 70.2 (+/- 14.2)).

Outpatient samples (SHC and Duke) - In addition, we also created separate outpatient dataset for SHC - 100 consecutive patients (67 women, mean (SD) age, 57.74(+/- 19.87)) and Duke - 101 consecutive patients (59 women, mean (SD) age, 73.06(+/- 15.3)), and these cases are independent from the internal SHC and Duke hold-out dataset.

Proposed System:

The PE prediction approach is outlined by the following problem statement: “Given a new patient encounter and access to prior structured EMR data (vitals, demographics, labs, inpatient and outpatient medication, diagnoses) predict the risk of pulmonary embolism (PE)”. This framing lends itself to be treated as a probabilistic classification
problem. The proposed workflow parses raw EMR data arranged as a timeline to transform into feature vectors of use in training a machine learning model based on PE imaging outcomes. For each patient, we defined their observation window as the 12 months leading up to a given prediction date (24 window of CT exam). Within the observation window, we created a feature engineering pipeline that computes a vector representation of the EMR snapshot of each patient by considering the temporal sequence within the records. The designed pipeline parse five core components of EMR - (1) all diagnosis codes (except current encounter); (2) all inpatient and outpatient medications; (3) all laboratory data raw values; (4) all collected vital sign data; (5) all demographics. Given the complexity of the EMR data and the requirement of temporality preservation, we carefully designed a EMR feature engineering pipeline able to parse varying types of EMR simultaneously while also tolerant of sparse records (a common limitation).

As machine learner algorithm, we used the same input features and compared between a regularized regression methodology, ElasticNet, and a novel deep learning model (PE Neural model) – an Encoder network with ReLU activation and sigmoid outcome. To optimize the hyperparameters of ElasticNet (regularization rate) and the PE Neural network architecture (number of hidden layers, learning rate, activation function, optimizer, number of epochs, dropout rate), we used Grid search on 10% training SHC training data to reduce overfitting. Trained models are then tested on hold-out intra- and new extra- institutional patient data as test sets as well as tested on separate intra- and extra- institutional outpatient population. We also compared the performance of trained machine learning models against three popular clinical scoring systems for PE- Wells, PERC, and revised Geneva.

**Results**

The performance of the machine learning models are summarized as AUC-ROC in Table 1. The models are only trained on a sub-sample of the SHC data in order to test the generalizability of the model on a same training population as well as on a different population from another institution. Both machine learning models scored high accuracy on the internal testset of 340 CT exams (PE neural 0.85 and ElasticNet 0.93) while ElasticNet model outperformed the PE Neural model (p = 0.013). Both machine learning models’ performance dropped on the external duke dataset compared to the internal hold-out SHC testset (PE neural 0.72 and ElasticNet 0.7). However, the AUC-ROC score stayed > 0.7 and both models performed equally well on the external data (p = 0.165) which shows the fact that even when trained on the SHC patients the models are generalizable to the Duke patients (Inpatient and Outpatient) population.

**Table 1: Quantitative analysis of the model’s performance - measured in-terms of AUC-ROC score**

<table>
<thead>
<tr>
<th></th>
<th>AUC-ROC on SHC data</th>
<th>p-value</th>
<th>AUC-ROC on Duke data (external testset)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hold-out testing on the internal SHC dataset and external Duke dataset (inpatient and outpatient)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ElasticNet model</td>
<td>0.93</td>
<td>0.0132</td>
<td>0.7</td>
<td>0.165</td>
</tr>
<tr>
<td>PE Neural model</td>
<td>0.85</td>
<td></td>
<td>0.72</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Comparison with clinical scoring systems on outpatients from the internal SHC and external Duke dataset</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machine learning models</strong></td>
</tr>
<tr>
<td>ElasticNet model</td>
</tr>
<tr>
<td>PE Neural model</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Clinical scoring</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wells score</td>
</tr>
<tr>
<td>PERC Score</td>
</tr>
<tr>
<td>rGeneva Score</td>
</tr>
</tbody>
</table>

Given the criteria for usability of the clinical scoring systems for computing pretest probability, we randomly selected 100 outpatient samples from SHC and 100 from Duke, and created another hold-out cohort. We used the models that were trained using SHC patients and tested it on ED patients separately to judge the model performance of the ED cases in parallel with three popular clinical scorings (Table 1). The PE Neural model performed significantly better than the all other models/criteria on the Stanford and Duke hold-out ED patients including the ElasticNet model on the Duke data (p = 0.01).

**Conclusion**

In conclusion, we found that achieving prediction models based on available retrospective structured EMR data can consider multitudes of patient-specific risk factors and dependencies in order to arrive at a PE likelihood recommendation model is possible and these models may be more accurately generalized to new population distributions. Future work is needed in investigating the ideal application of these prediction models for clinical imaging decision support systems in suspected PE and ultimate effect on imaging utilization.
Implementing Predictive Analytic Models in Diverse Healthcare Systems: A Qualitative Study with Operational, Informatics, and Front-line Personnel

Natalie C. Benda, PhD¹, Erika Abramson, MD, MS¹, Lala Tanmoy Das, MS¹, Katherine Blackburn, MSPH², Amy Thoman, BS¹, Rainu Kaushal, MD, MPH¹, Jessica S. Ancker, PhD, MPH¹

¹Weill Cornell Medicine, New York, NY; ²University of Florida, Gainesville, FL

Introduction

Predictive analytics has the potential to identify vulnerable patients in time to prevent poor outcomes or allocate valuable resources, but in practice, the expectations have not been realized to the extent anticipated. Those attempting to implement predictive algorithms in medicine have encountered challenges related to: making the algorithms transparent to end users, ensuring algorithms are clinically meaningful, and establishing generalizability and interoperability between sites.¹,² Due to these issues, many models validated in research contexts are not meaningfully utilized in a clinical environment.³

The present study utilizes data analytic techniques and key informant interviews to develop a predictive algorithm to proactively identify patients who have high utilization of the healthcare system, so that preventative care interventions may be implemented earlier. Interviewees include operational, informatics, and front-line personnel to gain feedback from key informants with various scopes of practice. The combination of methods addresses gaps in current research that there is insufficient stakeholder engagement early on in algorithm development and the discrepancies between algorithms that are statistically predictive but not clinically meaningful or actionable.

We will focus on the results of the key informant interviews which aim to determine which stakeholders would find the predictive algorithm most useful, under what contexts, and how the algorithm may be operationalized across three diverse healthcare organizations. The results will be discussed in the broader context of important considerations in gaining early feedback from stakeholders for the development of clinically meaningful, actionable predictive algorithms.

Methods

We utilized semi-structured interviews to elicit feedback from key stakeholders including: operational leadership (e.g. chief medical officers, chief strategy officers), informatics personnel, and potential end users (e.g. care managers and primary care providers). Interviewees were recruited from healthcare institutions affiliated with the Clinical Data Research Network (CDRN) in New York City (INSIGHT) and the CDRN in Florida (OneFlorida). CDRNs are partnerships organized through the National Patient-Centered Clinical Research Network (PCORnet) to enable clinical research that is faster, easier, less costly and more relevant to patients’ needs.

Interview questions were developed based on Sittig and Singh’s sociotechnical model for studying health information technology in complex adaptive healthcare systems.⁴ Interviews covered broad themes, including: the utility of predictive modeling for the given context as well as barriers and facilitators to implementing and using predictive models in an applied environment. All interviews were conducted by the same moderator (NCB) accompanied by at least one additional member of the analysis team (EA, JSA, KB, TD). Interviews were audio recorded and transcribed. At least two members of the analysis team inductively coded each interview using group consensus coding.⁵ The institutional review board at Weill Cornell Medicine determined that this study was not human subjects research.

Results

Forty-eight total participants were interviewed to thematic saturation. Table 1 describes emerging themes with supporting quotations from interviewees.

<table>
<thead>
<tr>
<th>Emerging Category</th>
<th>Emerging Theme</th>
<th>Exemplar Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational resources</td>
<td>The predictive model is irrelevant if there is no infrastructure (process, resources, and band-with) to support the necessary subsequent actions</td>
<td>“The worst-case scenario would be you tell me I have a high-risk patient for whom I should be intervening on, but I’m not armed with any additional interventions... Then it feels probably just bad to say, ‘Hey, they’re really sick.’ And you’re like, ‘I know that!’” (End user) “We have to face into the reality institutionally of; if we’re going to surface this information, are there going to be resources to actually do something about it?” (Informatics)</td>
</tr>
<tr>
<td>Technical challenges</td>
<td>Credibility and perceptions</td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------------</td>
<td>---</td>
</tr>
<tr>
<td>Different sources of data have different cost-benefit tradeoffs</td>
<td>It is important to differentiate from the multitude of predictive models being developed</td>
<td></td>
</tr>
<tr>
<td>Local customizations of electronic health record systems</td>
<td>Arranging people and processes is more challenging than technical aspects of implementation</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Key themes and higher level categories identified to date. Following each quote, the role of the person who conveyed this information (i.e. operational, informatics, or end user personnel) is listed in parentheses.

Discussion
Predictive analytics has the potential to aggregate complex information and inform key decisions in a healthcare setting. While creation of these models is data-driven, it is the humans in the system who must take action on the information to implement steps to facilitate patient care. Therefore, it is critical to understand practitioner needs, so that the models delivered may be meaningfully implemented and utilized. The results of this work demonstrate that successful integration of predictive models depend on a diverse set of factors related to institutional infrastructure, hardware/software constraints, the model itself, and the needs of front-line personnel. The results of this study will be discussed in terms of an implementation science framework to provide considerations for implementing predictive analytic models into applied healthcare settings.

Acknowledgements
This work was funded by the Patient Centered Outcomes Research Institute (HSB-1604-35187; PI, Kaushal).

References
Detecting Seasonal, Holiday, and Rare Events from Trauma Data in the Electronic Health Record

Timothy Bergquist, BS¹, Vikas Pejaver, PhD¹, Noah Hammarlund, PhD¹, Sean D. Mooney, PhD¹, Stephen Mooney, PhD¹
¹University of Washington, Seattle, WA

Introduction

Adoption of Electronic Health Records (EHR) has increased dramatically over the past two decades, creating opportunities to ask interesting population health data science questions. Most studies have focused on more traditional population health questions such as tracking chronic and infectious diseases and the geospatial mapping of those diseases.¹ The variety of patients that come to hospitals gives researchers opportunities to explore non-traditional population health questions. Recently, Melamed et al. linked diagnosis codes to seasonal trends and discovered a novel seasonal disease while also confirming previous literature on seasonally linked diseases.²

We explore the ability of EHRs to be used in detecting expected and unexpected trends as well as single-day injury or poison events at the university’s emergency departments. Our analysis finds codes that are statistically “enriched” (overrepresented) seasonally and daily. For instance, we confirm that injuries from snow sports like skiing, snowboarding, and snowmobiling are enriched in the winter months. Our work thus showcases the novel opportunities that EHR data can bring to population health.

Methods

We gathered ICD-9-CM diagnosis codes from all visits to the University of Washington Medicine system where at least one code fell in the category of Injury and Poisoning (E000-E999 and 800-999). The visits numbered over 3,000,000 and spanned from 1993 to 2017 across three main hospitals. We looked for enriched diagnosis codes for each of the four seasons and for each of the 365 days of the year. Each code count was propagated to its parent node in the ICD-9-CM hierarchy to enable detection of high-level diagnosis codes that are not used for billing purposes. To test for enrichment, we established a baseline frequency for each diagnosis code by counting the number of times the diagnosis code appears versus the total number of codes in the timespan of interest.

To test for seasonal enrichment, we grouped the code counts into each of the four seasons and used a binomial test to compare each of the code counts in the seasons to the expected rate of appearance calculated from the other three seasons. To test for each of the 365 days of the year, we established a baseline frequency for each code using all the data across the 24 years and performed a binomial test for each code in each day, where the total number of codes in the day were trials and number of appearances for the code of interest where success. In this manner, we generated a p-value for each code and could calculate a score by taking the negative log of the p-value.

Results

Seasonal Enrichment

We found that winter and summer had the most interesting profiles of seasonal enrichments. Winter was enriched for codes relating to winter sports, cold weather-related injuries, and slips and falls. Interestingly, toxic effect of alcohol as well as complications from surgical procedures were enriched. The summer months were enriched for injuries from animal and insects, burns and firework accidents (4th of July), boating accidents, and motorcycle accidents.
Figure 1. Visualization of the ICD-9-CM enriched codes for the summer season (June, July, August). The size of each node is proportional to the significance of the association between that code and summer. The lowest level nodes are the leaf or billable codes in the ICD-9-CM hierarchy, while the parent nodes are the increasingly broader categories of diagnosis. The most enriched codes are injuries related to fair weather activities such as boating accidents, fireworks from 4th of July, bug bites, and motorcycle accidents.

Day Enrichment

We identified 100 days on which at least one code was still enriched after a Bonferroni correction. Looking at the top scoring days with the highest scoring codes, we find that holidays and identifiable environmental events are disproportionately represented.

Days like July 4th and 5th have the expected enrichment of firework related accidents. January 1st has an enrichment of brawls, poisoning due to alcohol, and injury from fireworks. December 24th has an enrichment of alcohol poisoning and injuries undetermined whether purposeful or accidental.

We also detected one-time events, where a single day in a single year had an unusually high number of codes. We detected the Nisqually earthquake on February 28, 2001, where a high number of people came into the clinic for earthquake related injuries (E909.0). We also saw an increase in the number of carbon monoxide poisonings (E868.3) on December 15, 2006, which was the Hanukkah Eve windstorm where people were barbequing in their homes.

Discussion

While the events detected in this study may be obvious (July 4th and fireworks), future use of this method could expand to non-trauma diagnosis codes and incorporate other, EHR specific, data such as notes or lab results to bring more context to the diagnosis code enrichments and better inform the appropriate public health response.

Conclusion

We show that we can detect seasonal, holiday, and rare events in the EHR. These results showcase the potential utility of EHRs for public health, given the scope and depth of EHR data. Additionally, public health officials may discover novel events to help prioritize public health policy if a similar analysis is applied to non-traumatic diagnosis codes.

References


Continuum of Interoperability in Oncology EHR Implementations

Elmer V. Bernstam, MD, MSE1 Jeremy Warner, MD, MS2, John C. Krauss, MD3, Edward Ambinder, MD4, Wendy Rubinstein, MD, PhD5, George Komatsoulis, PhD5, Robert S. Miller, MD5, James L. Chen, MD6

1School of Biomedical Informatics and Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, TX; 2Vanderbilt University Medical Center, Nashville, TN; 3University of Michigan Medical School, Ann Arbor, MI; 4The Tisch Cancer Institute at the Icahn School of Medicine at Mount Sinai, New York, NY; 5CancerLinQ® LLC, American Society of Clinical Oncology, Alexandria, VA; 6Division of Medical Oncology and Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio.

Introduction

Interoperability is the “degree to which two or more systems, products or components can exchange information and use the information that has been exchanged.”1 With the widespread implementation of electronic health records (EHRs), there has been a great deal of attention focused on the ability to share structured (clinical) data between institutions; or, to be more precise, between EHR implementations2,3. However, interoperability has often been discussed as a binary property without consideration of the purpose. Clearly, it is possible for two systems to be able to share some data (e.g., demographics), but not others (e.g., laboratory results). In addition, the sharing of a particular data type may be suitable for one purpose but not another. Thus, interoperability is a continuum, rather than an absolute. To quantitate EHR interoperability, we determined the standards in use at oncology practices.

The American Society of Clinical Oncology (ASCO) launched the CancerLinQ platform that collects, aggregates, and transforms EHR data from oncology practices, for quality improvement and discovery.4 This presents the opportunity to evaluate the standards in use for structured data across a sample of oncology practices as well as the degree to which the EHRs implemented at these practices are interoperable with respect to structured data.

Methods

CancerLinQ data from 47 practices were analyzed with respect to the standards and variability of structured data including race, diagnoses, encounters, cancer staging, selected cancer-relevant medications, lab values and biomarkers. De-identified and aggregated data are presented. EHRs represented included ARIA, MOSAIQ, Allscripts, Centricity, Epic Clarity, Intellidose, NextGen, and OncoEMR. Although CancerLinQ collects individual patient data, this analysis was performed on aggregated data by vendor system. Our analysis focused on the data representation, rather than the content (e.g., the standard for representing laboratory results, rather than the laboratory results themselves); thus, no patient data were shared and/or analyzed.

We also analyzed a separate de-identified, aggregated data set from 48 installations of one of five EHR vendor products (Allscripts, ARIA, Epic Clarity, MOSAIQ, and OncoEMR) grouped by de-identified vendor. The specific data types analyzed are listed in Table 1 (results). For each data type, we divided the number of distinct inbound versions of a value mapped to a single codified value (e.g., the ways in which a particular laboratory test result was represented), divided by the number of implementations of that EHR. In addition, we calculated the total number of representations across all 5 of the EHRs and divided by the total number of installations of all five EHRs to assess the mean number of variants across all EHRs.

Results

Only 6 practices (13%) used non-standard race representation. All practices used ICD-9/10 for diagnoses. There was variability in coding of encounters. Sixteen practices always used Current Procedural Terminology (CPT), 5 practices always used Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) and 26 practices used multiple standards. Multiple cancer staging systems were used. An average of 48% (range 11%-105%; including patients staged more than once) of patient records included coded staging information. Only one practice used a standard (Logical Observation Identifiers Names and Codes; LOINC) for laboratory data. No standards were used for medications or biomarkers. Table 1 shows the number of distinct inbound names for selected lab tests, medications, and biomarkers across implementations for the five EHRs with sufficient numbers of installations to make assessments. Each cell

1304
the table represents the number of distinct ways that the data item was represented in our sample, divided by the number of implementations. The row “All” represents the total number of distinct ways that the data was represented across all 5 system types. The row “Mean” represents the total number of inbound values for that data element (i.e. ‘All’) divided by the total number of installations of all EHRs.

### Table 1: Number of distinct names for selected data types

| Data Type | ALT | BUN | Creat | Hct | Hgb | Hct | Neutrop | Neutrop | Neutrop | Neutrop | WBC | ALT | BUN | Creat | Hct | Hgb | Hct | Neutrop | Neutrop | Neutrop | Neutrop | WBC |
|-----------|-----|-----|-------|-----|-----|-----|---------|---------|---------|---------|-----|-----|-----|-------|-----|-----|-----|---------|---------|---------|---------|-----|-----|-----|-------|-----|-----|-----|---------|---------|---------|---------|-----|-----|-----|-------|-----|-----|-----|
| A         | 0.3 | 0.2 | 0.3   | 0.5 | 0.7 | 1.3 | 0.8     | 0.3     | 0.7     | 0.3     | 0.7 | 0.7 | 0.5 | 5.4   | 1.6 | 0.4 | 0.8 | 2.6     | 0       | 0       | 0.1     | 0.3 | 0.3 | 0.1 | 0.1     | 0.3 | 0.1 | 0.1 |
| B         | 1.0 | 0.6 | 1.4   | 1.0 | 1.2 | 1.8 | 1.2     | 1.8     | 1.6     | 3.6     | 5.2 | 2.4 | 2.0 | 0.6   | 0.6 | 3.2 | 2.2 | 0.2     | 0.6     | 0.6     | 1.2     | 0.6 | 1.2 | 0.6 | 0.6     | 0.6 | 1.2 | 0.6 |
| C         | 1.8 | 1.8 | 1.3   | 2.0 | 6.8 | 2.0 | 2.5     | 3.0     | 6.5     | 2.3     | 3.3 | 12.3| 2.8  | 1.3   | 2.3 | 2.5 | 0     | 0.3     | 0.3     | 1.5     | 0.3 | 0.3 | 1.5 | 0.3     | 0.3 | 1.5 | 0.3 |
| D         | 1.3 | 1.0 | 1.0   | 1.3 | 4.3 | 3.0 | 2.7     | 2.7     | 0.7     | 1.7     | 0.3 | 6.7  | 1.7  | 0.0   | 1.7 | 2.0 | 0     | 1.0     | 0.3     | 0.7     | 1.0 | 0.3 | 0.7 | 1.0     | 0.3 | 0.7 | 1.0 |
| E         | 1.8 | 1.9 | 1.8   | 2.2 | 3.2 | 3.3 | 3.0     | 3.2     | 0.5     | 1.2     | 0.4 | 1.4  | 0.7  | 0.0   | 0.9 | 0.9 | 0     | 0.3     | 0.8     | 0.3     | 0.2 | 0.3 | 0.2 | 0.3     | 0.8 | 0.3 | 0.2 |
| All       | 37  | 35  | 37    | 46  | 72  | 81  | 69     | 60      | 53      | 58      | 53  | 239 | 68   | 18    | 56 | 95  | 1     | 8       | 20      | 17      | 7     | 8     | 20    | 17      | 8     |
| Mean      | 0.8 | 0.7 | 0.8   | 1.0 | 1.5 | 1.7 | 1.4     | 1.3     | 1.1     | 1.2     | 1.1 | 1.5 | 1.4  | 0.4   | 1.2 | 2.0 | 0     | 0.2     | 0.4     | 0.4     | 0.1  | 0.2 | 0.4 | 0.4     | 0.4 | 0.4 | 0.4 |

### Discussion

In this de-identified sample of 47 oncology practices, standards were used consistently for administrative data (diagnoses and encounters), often for race, and rarely for clinical data including medications, laboratory tests, and biomarkers. For data types where no standard was used, there was heterogeneity among vendor implementations in the consistency of representation. For example, vendor A typically has a mean number of inbound values less than 1.0 (reflecting homogeneity of values across implementations) but has a large mean (5.4) for inbound values of dexamethasone. The high mean number of inbound values seen for medications, particularly for specific vendors (e.g., vendors B and C), raises the question of whether standardization of EHR implementations might reduce confusion arising from medication spelling or trade names and have the potential to reduce medication errors.

We reviewed a relatively large number of practices with respect to common general and oncology-specific data types, including biomarkers. Other specialties may have different patterns of EHR implementation. Further, practices who choose not to participate in the CancerLinQ program may differ from the practices in our sample. Finally, our analysis does not address unstructured (text) data that pose different interoperability challenges compared to structured data.

Our results suggest that interoperability with respect to structured data is largely limited to billing data; probably due to payer mandates. Multiple projects, including CancerLinQ, have been able to map structured data across EHR implementations, involving significant manual effort. Large national and international efforts have been able to use observational data to reach novel and interesting conclusions. However, our results suggest that for the foreseeable future, interoperability will continue to require significant manual effort.

Institutions make EHR purchasing decisions in part on the basis of perceived interoperability. For example, an outpatient practice may choose the same vendor as a hospital with which they are affiliated. Our next steps will include analyses to determine whether it is the vendor or the implementation that has the most influence on interoperability. In other words, if two institutions share the same vendor, are their implementations more likely to be interoperable?

### Conclusion

With the exception of administrative data, few oncology practices used standards for structured data through their EHR. There were differences by vendor with respect to the likelihood of being able to share data of a specific type with other implementations of that vendor’s product.

### References

Introduction

As the population ages, a key public health priority is identifying older individuals at higher risk of suffering an adverse health event. Identifying risk among the older adult population is a challenging task because chronological age is not a generalizable predictor of health status. Despite the utility of frailty to identify older adults at risk, outcome measures for frailty are not routinely used to predict risk in clinical settings, specifically on a population level.

The advent of electronic health records (EHRs) has facilitated the aggregation of many disparate pieces of clinical information collected about a given patient. In contrast to surveys, EHR data can be queried on a population level and integrated within a clinical setting making it well suited for automatic calculation of risk scores, including frailty. Indeed, EHRs represent a valuable source of data for frailty scoring, and ongoing development of frailty measures should consider what data is available in EHR to best use this growing data source.

In this study, we surveyed the literature to identify the most commonly-used variables by new and established measures of frailty. To characterize how well these variables are covered in EHR, we performed an overlap analysis between the variables uncovered in our literature review and the variables most commonly available in EHR systems (most commonly-available based on expert feedback and meaningful use criteria established by the Centers for Medicare & Medicaid Services).

Methods

We conducted a systematic review of literature to identify frailty instruments published between 1998 and March 2018. Frailty variables used in each of the frailty instruments were extracted, grouped, and categorized by type (e.g., clinical, social, behavioral, health services-related). The availability of the extracted frailty variables was evaluated by experts across various data sources (e.g., EHRs, administrative claims, and surveys).

Results

We identified 138 studies representing the same number of unique frailty instruments. We extracted 539 unique variables from the frailty instruments. Clinical variables were the most commonly-used variables across the frailty instruments (n=516; 87%), while social and health services factors included considerably fewer variables (n=33; ~5% and n=32; ~5%). Most frailty instruments require at least one variable that is not routinely available in EHRs or claims (n=115; ~83%). Only 23 frailty instruments have the potential to completely rely on EHR and/or claims data, and therefore possibly be operationalized on a population-level dataset (Table 1). We also analyzed the distribution of data types (e.g., EHRs vs. surveys) among variables extracted from the reviewed frailty instruments (Figure 1).

Table 1. Potential data sources of variables used in reviewed frailty instruments

<table>
<thead>
<tr>
<th>Common Data Source</th>
<th>Items(^b) N (%)</th>
<th>Papers(^c) N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Claims</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-clinical Data</td>
<td>20 (3.4)</td>
<td>26 (18.8)</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>3 (5)</td>
<td>9 (6.5)</td>
</tr>
<tr>
<td>Laboratory Tests</td>
<td>2 (3)</td>
<td>2 (1.4)</td>
</tr>
<tr>
<td>Non-clinical Data</td>
<td>13 (2.2)</td>
<td>12 (8.7)</td>
</tr>
<tr>
<td>Medications</td>
<td>1 (2)</td>
<td>1 (1.7)</td>
</tr>
<tr>
<td>Signs</td>
<td>105 (17.7)</td>
<td>81 (58.7)</td>
</tr>
<tr>
<td>Symptoms</td>
<td>49 (8.3)</td>
<td>77 (55.8)</td>
</tr>
</tbody>
</table>

1306
<table>
<thead>
<tr>
<th>EHR-Mixed</th>
<th>1 (2)</th>
<th>3 (2.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Non-clinical Data</td>
<td>1 (.2)</td>
<td>3 (2.2)</td>
</tr>
<tr>
<td>EHR-Structured</td>
<td>229 (38.6)</td>
<td>118 (85.5)</td>
</tr>
<tr>
<td>• Diagnoses</td>
<td>113 (19.1)</td>
<td>96 (69.6)</td>
</tr>
<tr>
<td>• Laboratory Tests</td>
<td>72 (12.1)</td>
<td>35 (25.4)</td>
</tr>
<tr>
<td>• Non-clinical Data</td>
<td>4 (.7)</td>
<td>6 (4.3)</td>
</tr>
<tr>
<td>• Medications</td>
<td>9 (1.5)</td>
<td>26 (18.8)</td>
</tr>
<tr>
<td>• Signs</td>
<td>26 (4.4)</td>
<td>77 (55.8)</td>
</tr>
<tr>
<td>• Symptoms</td>
<td>5 (.8)</td>
<td>4 (2.9)</td>
</tr>
<tr>
<td>Survey</td>
<td>170 (28.7)</td>
<td>115 (83.3)</td>
</tr>
<tr>
<td>• Non-clinical Data</td>
<td>39 (6.6)</td>
<td>65 (47.1)</td>
</tr>
<tr>
<td>• Signs</td>
<td>85 (14.3)</td>
<td>103 (74.6)</td>
</tr>
<tr>
<td>• Symptoms</td>
<td>46 (7.8)</td>
<td>74 (53.6)</td>
</tr>
<tr>
<td>Total Frailty Vars/Instruments</td>
<td>593 (100.0)</td>
<td>138 (100.0)</td>
</tr>
</tbody>
</table>

Figure 2: Distribution of potential data sources for 593 variables used in 138 frailty instruments

EHR-ST: EHR-Structured; and, EHR-FT: EHR-Free Text; Sorted based on the ratio of survey-only variables within each frailty instrument
Frailty instruments on the left predominantly include variables that are collected by surveys, while instruments on the right mostly include EHR-based variables. Frailty instruments in the middle can use a mix of data sources.

Discussion & Conclusion

Given the widespread adoption of EHRs and the push for value-based outcomes in the U.S., EHR-based frailty instruments can be a useful tool to stratify risk and better allocate resources among the older adult population. Unfortunately, many existing frailty measures are predominantly survey-based, are time-consuming to administer and fail to leverage the wealth of possible information in EHRs. More work is needed to develop and assess standardized, pragmatic, and generalizable EHR-based frailty instruments for clinical or population health management purposes.

References

Protecting Patient Privacy in Survival Analyses

Luca Bonomi, PhD1, Xiaojian Jiang PhD2, Lucila Ohno-Machado, PhD MD1
1UCSD Health Department of Biomedical Informatics, La Jolla, CA
2UT-Health School of Biomedical Informatics, Houston, TX

Introduction. In healthcare applications, survival analysis can provide useful insights on effectiveness of treatments, hypotheses testing, and outcome prediction through the analyses of temporal events.1,2 A variety of models have been proposed to conduct survival analyses, namely: parametric, semi-parametric, and non-parametric, which differ on the assumptions of the underlying distribution of time-to-events (e.g., recurrence, re-hospitalization, death). On one hand, parametric methods assume that the underlying data distribution is known a priori, and use statistical models (e.g., deep learning models) to learn the relationship between covariates and the outcome. On the other hand, non-parametric/semi-parametric methods require fewer assumptions and estimate the survival distribution directly from the data. While these methods can benefit clinical and epidemiologic research, sharing survival results may pose privacy concerns. Several previous privacy studies have shown that sharing aggregated results may lead to privacy disclosures.3,4 As survival analyses rely on statistical primitives (e.g., counts of time-to-events), they pose similar privacy risks. For example, an adversary who: (1) has knowledge of the time-to-events of individuals in the case and control group at a certain time (e.g., previously released survival curves), and (2) knows that a person of interest recently joined the study, may infer the presence of such an individual in the specific subgroup (e.g., in the “case” group of a case-control study), as the released curves are updated. Consider for example, the situation represented in Figure 1, where s2 is a previously released survival curve and s1 is a new updated version of the curve. Then, the difference between these curves reveals that a new patient with an event at time 42 recently joined the group, thus potentially revealing sensitive phenotypes (i.e., the individual is a “case”). Current privacy solutions for survival analysis provide limited privacy protection against such an attack. Several security solutions rely on cryptographic methods to encrypt data statistics to protect data access.5 However, those methods are vulnerable to privacy attacks in the presence of a malicious user who has access to the statistics. Other approaches use data perturbation to provide privacy protection.6 However, those solutions are unable to provide provable guarantees, as those methods may be vulnerable in the presence of an adversary with auxiliary information (e.g., previously released curve). In this paper, we outline our on-going study on privacy methods for survival analyses, which aims at assessing the privacy risks in survival analyses and developing privacy-protecting solutions based on the formal privacy model of differential privacy.7

Method. In this work, we focus on the study of the Kaplan-Meier non-parametric model, which is one of the most widely used methods in survival analyses.

**Privacy Solution.** In our work, we use the notion of \( \varepsilon \)-differential privacy,7 which provides provable privacy protection to the individuals contributing the data by introducing carefully calibrated noise. Intuitively, differential privacy ensures that the probability distribution on the released statistics should be “roughly the same” regardless the presence/absence of any individual, thus providing plausible deniability. Traditional differential privacy solutions introduce a perturbation noise on the output results. Because survival studies may span over a large period of time, traditional output perturbation methods may produce

![Figure 1 An adversary who compares the two curves s1 (updated curve) and s2 (past curve known to the adversary) can learn that their difference is due to an event at time unit 42. Thus, learning that an individual patient with an event at time 42 contributed to the updated curve.](image)

**Table 1** Mean absolute inference error (MAIE) in time units and mean absolute error (MAE) in probability for the original survival curve (KM) and our differentially private solution (DP-KM) with different privacy parameter (\( \varepsilon \)) and data size (N).

<table>
<thead>
<tr>
<th>Privacy Parameter (( \varepsilon ))</th>
<th>N=1000</th>
<th>N=10000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KM MAE</td>
<td>DP-KM MAE</td>
</tr>
<tr>
<td>0.5</td>
<td>1.75</td>
<td>222.65</td>
</tr>
<tr>
<td>1</td>
<td>1.75</td>
<td>218.91</td>
</tr>
<tr>
<td>2</td>
<td>1.75</td>
<td>218.91</td>
</tr>
<tr>
<td>4</td>
<td>1.75</td>
<td>231.26</td>
</tr>
<tr>
<td>8</td>
<td>1.75</td>
<td>221.76</td>
</tr>
</tbody>
</table>

1308
survival curves with limited usability. To overcome this challenge, we propose a solution that reduces the number of data releases by compacting the stream of events into partitions, in which the survival probability can be accurately computed over time using an input perturbation mechanism. Overall, our framework comprises of three main steps: (1) data partitioning, (2) survival curve computation via input perturbation, and (3) post-processing to enforce consistency in the released curve (e.g., monotonicity).

Privacy Risk Measures. In the $\varepsilon$-differential privacy model, the adversary’s probability of inferring the contribution of an individual in the survival curve is only a factor $(1+\varepsilon)$ larger than the probability without the inclusion of such an individual in the study. Thus, the privacy parameter $\varepsilon$ bounds the privacy risk of disclosing the participation of each individual in the study. In addition, we consider an empirical inference risk related to the disclosure of exact time-to-events. In this case, the adversary has prior knowledge about the time-to-events of other individuals and, by observing the released curve, the adversary aims at learning the exact time-to-event of an unknown individual in the study. We evaluate this empirical privacy risk by measuring the mean absolute inference error (MAIE) in reconstructing the time-to-event. Intuitively, the inference error measures how accurately the adversary can infer the time-of-event of the target in the data. Higher values of inference error indicate lower privacy risk for the patients participating in the study.

Results. We conducted our preliminary evaluations using the Surveillance Epidemiology and End Results (SEER)$^8$ dataset from which we sampled $N \in \{1000, 10000\}$ patients from 707,157 breast cancer patients with first diagnosis from 1973 to 2015. We evaluated the privacy in terms of mean absolute inference error (MAIE), while the utility of the released curve by our differentially private Kaplan-Meier method (DP-KM) was measured in terms of mean absolute error (MAE) with respect to the original non-private Kaplan-Meier survival curve (KM). From Table 1, we observe that the MAE for the non-private curve is as low as 1.75, which indicates that a knowledgeable adversary can re-construct the time-to-event with high accuracy (i.e., within 2 time units). With our DP-KM method we can significantly lower such a privacy risk, in fact the inference error of the adversary is now around 200 time units. We also notice that the released curve by our method closely resembles the original survival curve, as we can see from the small MAE values in Table 1 and from the visual comparison in Figure 2.

Conclusion. Our preliminary results quantified the privacy risk in sharing exact survival results, in particular when the number of patients in the study is small. Furthermore, we demonstrated that our differentially private method can significantly reduce the privacy risk while retaining the utility of the survival curve. As a future research direction, we plan to extend our privacy analysis to other survival models.

Reference
A Pragmatic Multi-system Opioid Use Disorder Computable Phenotype for the Emergency Department

David Chartash, PhD1, Hyung Paek, MD, MSEE2, James D. Dziura, PhD, MPH3, Bill K. Ross4, Daniel P. Nogee, MD3, Eric Boccio, MD3, Cory Hines, MD5, Aaron M. Schott, MD5, Molly M. Jeffery, PhD6, Mehul D. Patel, PhD5, Timothy F. Platts-Mills, MD, MSc5, Osama M. Ahmed, BS3, Cynthia A. Brandt, MD, MPH1,3, Katherine C. Couturier, MD, MPH3, Edward R. Melnick, MD, MHS3

1 Yale Center for Medical Informatics, Yale University School of Medicine, New Haven, CT; 2 Yale New Haven Health, Information Technology Services, New Haven, CT; 3 Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT; 4 North Carolina Translational and Clinical Sciences Institute, University of North Carolina School of Medicine, Chapel Hill, NC; 5 Department of Emergency Medicine, University of North Carolina School of Medicine, Chapel Hill, NC; 6 Department of Emergency Medicine and Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Introduction

Deploying accurate computable phenotypes in pragmatic trials requires a trade-off between precise and clinically sensible variable selection. For individuals with Opioid Use Disorder (OUD) who would benefit from medication-assisted treatment (MAT) in the clinical setting of the emergency department this study developed a phenotype using clinically available variables, and validated it in two health systems. OUD is a “problematic pattern of opioid use leading to clinically significant impairment or distress”2. The study was to support a pragmatic trial of a clinical decision support system to aid emergency physicians evaluating patients with OUD by initiating MAT and referring them for follow-up care. In addition to validating the phenotype’s accuracy, the team performed an evaluation of the gold standard diagnostic criteria. For OUD, the gold standard is meeting 2 of the 11 criteria and further qualifications listed in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). These criteria generally describe “A problematic pattern of opioid use leading to clinically significant impairment or distress”2.

Methods

125 random patient charts were retrieved from Health System A, and 106 from Health System B. Both health systems consist of academic medical centers with extensions into urban and rural hospitals providing the majority of regional care. Not all charts met the computable phenotype. These charts represented a sample of the OUD patient-population, rather than the true prevalence of disease. Charts from Health System A were assessed using the algorithms in Figure 1, with two emergency medicine-trained physicians as reviewers and a third as adjudicator. In Health System B, the same validation process was performed; however, due to a small sample meeting algorithm 2, the algorithms were combined into a single phenotype. Given that population invariance conditions were not met, both in terms of the denominator problem3 and site comparison, a population sensitivity and specificity was calculated rather than a true prevalence-based sensitivity and specificity.4 Reviewers were asked to identify the DSM criteria associated with each case as a reference standard.

Results

In Health System A, inter-rater reliability was high, with a $\kappa$ of 0.9. Algorithm 1 had a positive predictive value of 0.96 (95% CI 0.863-0.995) and a negative predictive value of 0.98 (95% CI 0.893-0.999). Algorithm 2 had a positive predictive value of 0.8 (95% CI 0.593-0.932) and negative predictive value of 1.0 (one-sided 97.5% CI 0.863-1). In Health System B, inter-rater reliability was high, with a $\kappa$ of 0.85. The combined phenotype had a positive predictive value of 0.95 (95% CI 0.851-0.989) and negative predictive value of 0.92 (95% CI 0.807-0.978).

The most common DSM-5 criterion selected by the reviewers at both Health System A and Health System B was “opioids are often taken in larger amounts or over a longer period than was intended”. In Health System A, the least common criteria were those which express social dysfunction related to the use of opioids (including “recurrent opioid
Figure 1: Flow diagram for inclusion in the phenotype algorithms from left to right (Algorithm 1 and Algorithm 2). MOUD: Medication treating Opioid Use Disorder.

use resulting in a failure to fulfill major role obligations at work, school, or home” or “important social, occupational, or recreational activities are given up or reduced because of opioid use”). In Health System B, the least common criterion was “important social, occupational, or recreational activities are given up or reduced because of opioid use”.

Discussion

The phenotype performed well, with excellent positive and negative predictive values. For medical decision making, this provides a highly sensitive and specific means to classify patients such that they can be evaluated for the initiation of MAT. The reviewers’ evaluation of DSM-5 criteria suggest that diagnostic decision-making is being primarily made on the use of opioids over a threshold of temporality or magnitude of prescription. Furthermore, the patients’ social history in the context of OUD\(^5\)\(^6\) was not found to be present as a diagnostic criterion within the charts. While the DSM-5 diagnostic criteria are specific in identifying and defining OUD, they were noted upon debrief as having generally been inferred by reviewers from the patient’s history and limited documentation. It is suspected that this inference is due to the limitations of social history and emergency department documentation derived from the documentation focus of the Centers for Medicare and Medicaid Services guidelines. As such, in the pragmatic context, patients classified as having OUD will require a complete social history and psychiatric evaluation during the MAT referral process with results to be used both in terms of clinical operation and improving diagnostic precision.

References

The Process of Developing, Validating and Operationalizing a Personalized Machine Learning Algorithm for Clinical Decision Support: A Case Study

Sara K. Chokshi, DrPH, 1 Ji Chen, MS, Roshini Hegde, MA, Javier Gonzalez,1 Eduardo Iturrate,2 MD, Yindalon Aphiryanaphongs, MD, PhD,1 Devin M. Mann, MD, MS1,2
1NYU School of Medicine, New York, NY; 2NYU Langone Health

Introduction: Clinical decision support (CDS) alerts are designed to promote appropriate vaccinations and other evidence-based practices. Effectiveness of CDS alerts is blunted by the expansion of alerts in the EHR in recent years, leading to a growing epidemic of “alert fatigue” (a term describing a clinician’s desensitization to an overwhelming number and quality of safety alerts),1,2 and causing clinicians to ignore or respond inappropriately to alerts. Optimization of CDS tools has historically focused on manual approaches that are time consuming, difficult to maintain, and static, limiting their scalability. In their application of Signal Detection Theory to the study of CDS effectiveness, Ong and Coiera found both clinician performance and user experience are compromised as a result of a large quantity of alerts and poor alert reliability.3 These system-level factors created a high degree of system “noise,” detracting from the perceptibility of appropriate alert fires. Individual-level factors, including clinicians’ bias toward ignoring alerts and poor signal detection, combine with system-level factors to degrade CDS effectiveness and user experience.4 Machine learning (ML) is a powerful tool for identifying patterns in complex data by using past data to predict future performance. ML applied to EHR data, specifically, shows signs of promise as a tool for improving safety and quality of care. An EHR- integrated ML algorithm may be a powerful tool to increase the signal to noise ratio of CDS alerts and positively impact clinician interaction with alerts.

Objective: The objective of this paper is to describe the process of the development, validation, and implementation of a novel, ML-based CDS model (SmartCDS ML) that suppresses low value vaccination alerts.

Methods: We employ a case study approach to describe the process by which we developed, validated and operationalized a ML model to suppress shingles vaccinations alerts that have a high likelihood of being ignored. Using the shingles vaccination alert as our initial use case, we leveraged historical EHR interaction data (clicks) and patient/provider sociodemographic data to build and train a ML model that could predict the likelihood of provider interaction with the shingles vaccination alert [in submission].

Results: Model development (SmartCDS ML). The initial component of the process relates to the building of the alert suppression ML model. The initial building of the model involved historical data query to extract data related to the shingles vaccine alert, for which key data elements to be extracted from EHR databases were determined using a combination of descriptive analysis and clinical experience. After data cleaning, analysis was conducted, including examination of the historical changes in alert to determine the optimal time period from which to extract data for model training. Initial analysis examined average response rates for the alert as well as provider interaction with the alert for the purposes of determining an appropriate protocol for assigning one unique provider to each alert encounter. Large variation between clinicians in terms of frequency of interaction with the alert (0-92%), prompted our team to construct a variable for individual clinician activity history which was expanded to a short and a long term activity history variable; additionally, we included demographic variables for both clinician and patient in the model. Once the predictive variables were constructed, we then built and trained the whole model (SmartCDS ML), further refined the predictive variables and optimized model parameters over the following month.

Building operational capacity for SmartCDS ML implementation. To implement the SmartCDS ML model we built a new data architecture (Figure 1). The overall Smart CDS system is broken into three components: (1) The data extraction module which identifies (a) outpatient visits (or inpatients) for the next day with the intent to identify upcoming vaccine alerts to suppress, and (b) queries the EHR and extracts the variables required to run the ML module; (2) SmartCDS ML itself: the machine learning module built as described above; and (3) The suppression module, which leverages a series of application programming interface calls to the EHR to communicate which alerts to suppress. The data extraction module queries the EHR and extracts features then passed to the ML module. We worked with our enterprise IT team to determine the appropriate web service to “call,” created the rules necessary to appropriately respond to the data sent to that endpoint and, if appropriate, suppress the target alert (the shingles vaccination alert in this case).
Once built, we validated the SmartCDS system with the shingles CDS alert. Using data from the prior month, we suppressed shingles vaccination alerts if the likelihood of the provider interacting with the alert in that visit fell below a predefined threshold, using a conservative threshold of 50% suppression. Applying the model to live data extracted from our predefined views, we optimized the ML script, allowing for further feature engineering. From this process we generated model predictions (to suppress alert or not) which were saved the local database. The model is re-trained daily incorporating latest data changes to update the score threshold used to make decisions on upcoming hospital visits the next day. The SmartCDS system was designed to be modular and orthogonal with regard to call frequency, instrumentation, and configuration, allowing for easy adaptation to new environments. To ensure that the system is running properly from a safety and operational perspective, and to monitor process outcomes of interest, we developed a reporting dashboard.

**Discussion:** This paper describes the steps and considerations involved in the development and implementation of a ML model for suppressing low value alerts in the EHR and its application to the shingles vaccination. As predicted in our simulation, validation of the SmartCDS system demonstrated substantial reduction in shingles vaccine alerts at a limited vaccine ordering expense. The development of a robust reporting structure allows for the logging and monitoring of the system and its impact on clinical outcomes, which are necessary to ensure the stability and safety of the system. Future work, in line with the goals of our learning healthcare system, will involve adapting and applying the SmartCDS system to other vaccine alerts, further development of a reporting dashboard with effective, user-centered data displays, and the inclusion of an implementation theory informed process for establishing institutionally-aligned thresholds for alert suppression.

**Conclusion:** A ML model can predict future alert interactions and safely suppress low yield alerts. These results support a full-scale implementation of the SmartCDS system to reduce alert fatigue and optimize vaccination ordering by reducing the amount of alerts required to facilitate clinicians’ vaccine ordering.

**References**
A bring-your-own-device (BYOD) model for contributing mobile health (mHealth) data enables real-world data collection as patients go about their daily activities. To date, most mHealth research studies provision a specific wearable device (i.e., Fitbit) and have a constrained study period during which data is collected. A BYOD mHealth model allows for capturing data from patients’ routine lives and has efficiencies at scale, allowing researchers to better understand patient trajectories in a real-world deployment of devices. There are growing examples of BYOD data contribution for the purposes of research including the PCORnet Inflammatory Bowel Diseases (IBD) Partners (formerly Crohn’s and Colitis Foundation of America) patient-powered research network and NIH’s All of Us Research Program, where participants can currently contribute their Fitbit data.

IBD Partners allows for a wide range of wearable devices and apps to be connected to our research platform. This facilitates mHealth data contribution for those who participate in our longitudinal Internet cohort study. These mHealth contributors can connect different devices over time and can view their data trends in the IBD Partners patient portal. Participants also contribute self-reported survey data on health outcomes, such as disease activity, as well as on patient-reported outcomes such as depression and anxiety.

While BYOD has many benefits, there are also challenges due to the diversity of both devices/apps and usage patterns that come with real-world data generation. As this BYOD data contribution model is still an emerging one, there is little known about how many patients will choose to contribute their mHealth data, and how those patients may differ from those who do not. We examine which brands are represented across mHealth contributors within the cohort, the patterns for device wear-time (usage) among these participants, and blocks of missingness where devices/apps were not used. In this oral abstract, we present an overview of the characteristics of a BYOD mHealth study, Precision VISSTA, which is an NIH-funded study that seeks to develop preprocessing, machine learning, and data visualization methods for mHealth data to generate precision health recommendations for patients with IBDs as the initial use case.

Methods

Descriptive statistics were used to summarize sample characteristics and to examine patterns across mHealth data for physical activity and sleep. Two-sample t-tests were used to evaluate differences between mHealth data contributors versus those who were not for disease activity (Simple Colitis Activity Index (SCAI) for ulcerative colitis; Simplified Crohn’s Diseases Activity Index (SCDAI) for Crohn’s diseases) and patient-reported outcomes for depression, anxiety, sleep disturbance, social relationships, and pain interference.

Results

Of the 10,090 patients enrolled thus far, 437 patients have contributed mHealth data. ~71% of patients who contributed mHealth data were female, 91% were White, the majority were between the ages of 18-60 years old (18.5% 18-25 years, 44.6% 26-40 years, 30.9% 40-60 years), and 78.3% had a college or graduate school education. Of those who connected device(s)/apps to contribute mHealth data, 62% had Crohn’s disease and 36% had ulcerative colitis. For distribution of device/app brands across the cohort, 67% were Fitbit, 10.3% Garmin, 5.6% Strava, 5.2% Under Armour, 4.1% Jawbone, 2.1% Moves App, 1.8% RunKeeper, 1.3% Withings, 0.8% Microsoft, 0.6% Map My Fitness, 0.6% Misfit, 0.4% Fatsecret, and 0.2% Nike+.

Figure 1a visualizes various patterns of step count data in two patients as an example; blocks of missing data are shown in white. Patterns differ across the day of the week for some participants, while others have more heterogeneity in their step count patterns. The majority of participants have contributed between 6 months to 3 years of mHealth data (Figure 1b) and the size of blocks of missing data ranged from one day to ≥4 weeks (Figure 1c).

On average, participants had 5,384 total steps per day and 6.8 hours of total sleep per night. Approximately 29% of who contributed mHealth data had at least 150 minutes of moderate-to-vigorous physical activity per week. We
observed differences in outcomes for mean IBD disease activity scores for those who contributed mHealth data versus those who did not (SCAI=2.9 vs. 3.1, p=0.003; SCDAI=124.7 vs. 140.7, p<0.0001). For mean patient-reported outcome scores in the domains of pain interference, depression, anxiety, fatigue, social relationships, those who contributed mHealth data had better scores than those who did not [49.75 vs. 51.40, p<0.0001 for pain interference; 49.48 vs. 50.23, p<0.0001 for depression; 51.12 vs. 51.99, p<0.0001 for anxiety; 53.84 vs. 54.30, p=0.01 for fatigue; 50.85 vs. 49.24, p<0.0001 for social relationships (higher score is better for social relationships; lower score is better for pain interference, depression, anxiety, and fatigue)].

Figure 1. Example of step patterns and distribution of data contributions and missing data block size.

When mHealth data is examined in aggregate for an individual and for the cohort, there are both common patterns (such as weekday vs. weekend patterns in sleep; see Figure 1a) and systematic differences between brands (Figure 2a and 2b). For some sleep variables, such as deep sleep, changes in the algorithms that process sleep data can also be seen (i.e., August 2017 in Figure 2b).

Figure 2. Patterns of sleep across brands.

Conclusions

Thus far, ~4% of the IBD Partners Internet cohort have contributed mHealth data using our BYOD model across various brands. Participants were mostly female, educated, and less than 60 years old. The majority have either Fitbit or Garmin devices. Those who track and contribute mHealth data have significantly better disease activity scores and patient-reported outcomes. However, the difference in scores may not all be clinically significant. When examined in aggregate across the population, there are systematic differences across brands of devices/apps for physical activity and sleep. Our analyses suggest that simple visualization techniques can expose important features of mHealth data, such as changes in software algorithms that process the data. Ongoing work seeks to address challenges in preprocessing to deal with missingness and sparsity and the batch effects seen across brands of devices/apps.

References

Precision VISSTA: Machine Learning Prediction and Inference for Bring-Your-Own-Device (BYOD) mHealth Data

Tim Coleman¹, Lucas Menth, PhD¹, Kimberly Glass, PhD²-³, David Gotz, PhD⁴, Nils Gehlenborg, PhD³, Arlene E. Chung, MD MHA MMCI⁴
¹University of Pittsburgh, Pittsburgh, PA, USA; ²Brigham & Women’s Hospital, Boston, MA, USA; ³Harvard Medical School, Boston, MA, USA; ⁴University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Introduction

Precision VISSTA is a bring-your-own-device (BYOD) mobile health (mHealth) patient-powered research study focused on Inflammatory Bowel Diseases (IBDs). Participants report longitudinal survey data on outcomes such as disease activity along with patient-reported outcomes like as sleep disturbance, while also contributing mHealth lifestyle data from various wearable devices and apps (24 types). IBD patients have extremely heterogeneous phenotypes with symptoms that fluctuate. Prior work has suggested an association between increased self-reported physical activity and decreased disease activity¹, while self-reported sleep disturbance has been associated with increased disease activity². However, the precise nature and quantity of activity and sleep associated with improved outcomes is not well established. Our mHealth dataset contains numerous features describing physical activity and a number of other lifestyle characteristics, allowing for large-scale analysis of the features most associated with IBD disease activity and symptoms. Because of the complex underlying relationships within the data, we considered a number of flexible machine learning (ML) approaches in order to avoid the rigid model structure imposed by most traditional statistical models. We leveraged recent theoretical results on inference for supervised learning ensembles to develop and implement permutation-style hypothesis tests for feature significance on these otherwise “black-box” models. The primary study objectives were: (1) to formally establish the predictive relevance of mHealth features in forming more accurate predictive models than could be obtained with survey data alone, and (2) to infer which specific mHealth lifestyle features are most predictive of outcomes for patients with IBDs.

Methods

To determine whether disease activity [Simplified Crohn’s Disease Activity Index or SCDAI for Crohn’s disease (CD); Simple Colitis Activity Index or SCAI for ulcerative colitis (UC)] and patient-reported outcomes (depression, anxiety, pain interference, sleep disturbance, social relationships) could be accurately predicted using mHealth lifestyle data, we constructed various machine learning models and conducted significance tests for features within this framework to account for potentially complex, nonlinear relationships. Our analysis takes data from three sources: (1) survey data on outcomes, (2) demographic information from surveys, and (3) mHealth data collected longitudinally as patients complete surveys, including device or app type. We aimed to determine whether mHealth data adds predictive power beyond what can be learned from survey data alone (data sources 1 and 2). For each outcome, we conducted a cross-validation (CV) analysis to select the optimal ML method for forecasting patient outcomes on their next survey. Models considered included best subset regression (Best Reg), elastic net, random forests (RF), conditional inference random forests (C-RF), gradient boosting (GrBoosting), partial least squares (PLS), and multivariate adaptive regression splines (MARS), with each model carefully tuned using the R package ‘caret.’ For each outcome, models were ranked according to their 5-fold cross-validated mean squared error (MSE) with better models attaining lower average rank in terms of generalization error. As a baseline, partial F-tests were conducted for each mHealth data type and outcome. We then conducted inference on those models identified as highly accurate in the CV analysis and tested whether a model trained on all features (i.e. demographic and prior survey data with mHealth data) attains equal generalization error to one trained using only prior survey outcomes and demographic information. Features with missing values were imputed using a RF model to iteratively predict missing values based on other complete feature information. The hypothesis tests for significance were carried out using a permutation-style procedure informed by recent work on inference with random forests.⁴⁻⁵

Results

The analytic sample had 539 inter-survey observations from 371 unique patients, of which 332 observations came from CD patients and 228 from UC patients. The cross-validation analysis consistently suggested that C-RF and elastic net models generated the most accurate predictions across outcomes. The overall tests for significance
indicate that, in aggregate, mHealth data was predictive of pain interference across all three tests with more modest evidence for an effect on SCDAI disease activity (CD), social relationship, and depression scores (Table 1). The more granular tests for predictive significance of individual mHealth features suggest that time spent in moderate-to-vigorous activity (active_duration) was predictive of pain interference and disease activity for patients with either CD (SCDAI) and UC (SCAI) in the elastic net models (Figure 1). The tests on the elastic net model also suggested that distance traveled throughout the day was predictive of disease activity for UC patients and also predictive of sleep disturbance, fatigue, and depression scores across patients. While total hours of sleep (total_sleep) was only predictive of disease activity for patients with UC (SCAI), it was also predictive for depression and pain. Total steps per day was only strongly predictive for disease activity in patients with CD. Tests conducted on the conditional inference random forest model detected fewer significant results, and water consumption was the only mHealth feature consistently shown to improve predictions across outcomes. We suspect, however, that this may be an artifact of its low correlation with other features rather than an indication of strong predictive power.

<table>
<thead>
<tr>
<th>Test\Outcome</th>
<th>Anxiety</th>
<th>Depression</th>
<th>Fatigue</th>
<th>Sleep</th>
<th>Social</th>
<th>SCAI</th>
<th>SCDAI</th>
<th>Pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>F Test</td>
<td>0.0498*</td>
<td>0.5428</td>
<td>0.0374*</td>
<td>0.5114</td>
<td>0.1571</td>
<td>0.3969</td>
<td>0.0045**</td>
<td>0.0017**</td>
</tr>
<tr>
<td>C-Forest Test</td>
<td>0.2298</td>
<td>0.6913</td>
<td>0.2138</td>
<td>0.6993</td>
<td>0.005**</td>
<td>0.8581</td>
<td>0.6124</td>
<td>0.001**</td>
</tr>
<tr>
<td>Elastic Net Test</td>
<td>0.8991</td>
<td>0.008***</td>
<td>0.4256</td>
<td>0.6663</td>
<td>0.0689</td>
<td>0.0889</td>
<td>0.2877</td>
<td>0.03*</td>
</tr>
</tbody>
</table>

Table 1. P-values for overall effect of aggregate mHealth features by outcome and test procedure. *p=0.05, **p=0.01

Figure 1. Conditional inference random forest and bagged elastic net results by mHealth data type and outcome.

Conclusion

Elastic Net and C-RF were identified as the top performing ML models with improved predictive accuracy when prior outcomes were included as features. C-RF models are inherently more flexible, and thus, we found fewer marginally significant predictive features since the models can better account for lost signal by detecting more complex signals from other available features. Best subset regression models performed poorly, demonstrating that traditional linear models fail to capture substantial predictive information contained within the mHealth data. Significance tests suggested that mHealth data are predictive of pain interference, even when accounting for prior patient-reported outcomes data. Moderate to vigorous physical activity (active_duration) was also highly predictive of IBD disease activity in both UC and CD patients in the elastic net models, while steps were predictive only for those with CD.

References

Quality Analysis of the All of Us Research Program Health Surveys

Robert M. Cronin, MD,1 Sarah Feng, MS,1 Brandy Mapes, MLIS,1 Roxana Loperena-Cortes, PhD,1 Regina Andrade, MS,1 David Schlundt, PhD,1 Kenneth Wallston, PhD,1 Mick P. Couper, PhD,2 Scott Sutherland, BFA,3 Cindy Chen, PhD,1 Joshua C. Denny, MD, MS1

1 Vanderbilt University, Nashville, TN, USA, 2 University of Michigan, Ann Arbor, MI, USA, 3 Vibrant Health, Fairfax, VA, USA

Introduction

The All of Us Research Program (All of Us) is recruiting a million participants reflecting the rich diversity of the U.S. population to advance the science of precision medicine[1 2]. A variety of data sources will contribute information about a participant including health surveys, electronic health records (EHRs), and biospecimens. Collecting these data into a central resource will enable deeper exploration of biological, clinical, social, environmental and behavioral determinants of health and disease. Health surveys have been an important part of observational research and will play an essential role in All of Us. Prior to public release, these materials were rigorously created, evaluated and approved[3]. Methods included creation of initial survey content, translation into Spanish, qualitative and quantitative assessment, and review by Program experts. There are three surveys made available to All of Us participants at baseline: Basics, Overall Health and Lifestyle. These surveys focus on demographic information, general health, alcohol consumption, substance use, and smoking.

All of Us began collecting health survey response data in May 2017. Participants answer health survey questions via an online participant portal application developed by the Participant Technology Systems Center (PTSC). The participant portal is available on both computers and mobile devices, and guides participants through surveys to be completed. The data are subsequently sent and stored in the Data and Research Center’s (DRC) Raw Data Repository (RDR). Standard integration processes are in place, including end-to-end testing, to ensure data is seamlessly sent from one system to the other. However, there is a need to ensure the utility and quality of the data is intact prior to its future release to researchers. The Quality Assurance assessment presented in this podium abstract will serve as the first instance of full health survey response data review since surveys were integrated into these complex systems and publicly released. The scope of this assessment was to extract, review, and analyze the All of Us survey responses to confirm data integrity, alignment with expected values from other studies (where appropriate), and data missingness.

Methods

We analyzed data from the first three surveys in All of Us received between May 2017 and December 2018. Survey questions are available at https://www.researchallofus.org/. All data were stored on a Federal Information Security Management Act (FISMA) approved virtual machine. Data was extracted and transformed from its structured organization within the RDR into a single file compatible with statistical software. A codebook, describing the survey hierarchy, served as a reference guide for this work. Identifiable string values such as name were excluded. Analyses of the data included frequency distributions, missingness, and flagging potentially problematic data elements.

Statistical analysis: Two major statistical software packages were used to conduct analyses for this project: SPSS and R. The majority of these analyses were conducted by a member of the DRC team with expertise in both quantitative analysis and psychosocial metrics. A second statistician on the DRC team provided additional review and replication of the analyses as a method for confirming the validity of the data and performance of the questions in instances when areas identified for more granular inspection were prioritized. For the majority of survey questions, a simple frequency distribution of selected responses was generated. To confirm the overall performance of the measures, we compared a selection of responses from the All of Us sample to those from other national cohort studies. In addition to frequency distributions of selected responses, frequency of missing data was also calculated for each of the survey questions based on the number of participants that should have seen the question as dictated by survey “skip” or “branching” logic, where some participants might have gotten certain survey questions and others would not.

Results

There are a total of 32 potential questions that a participant could answer in the Basics, 24 in Overall Health, and 31 in Lifestyle. 116,319 participants completed the Basics, 112,544 participants completed Overall Health, and 111,512 participants completed Lifestyle. Frequencies of the All of Us responses were similar to other national surveys, but there were a few differences (Table 1). While most questions had a less than 10% missing rate, some of the following questions had higher missing rates (Table 2).
Table 1: Comparison of selected All of Us survey responses to other national surveys

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
<th>All of Us %</th>
<th>National Survey: %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have you smoked at least 100 cigarettes in your entire life? (There are 20 cigarettes in a pack.)</td>
<td>Yes</td>
<td>39%</td>
<td>NHIS: 38% NSDUH: 57%</td>
</tr>
<tr>
<td>How often did you have a drink containing alcohol in the past year?</td>
<td>Never</td>
<td>18%</td>
<td>NSDUH: 35%</td>
</tr>
<tr>
<td>In your LIFETIME, which of the following substances have you ever used?</td>
<td>Marijuana</td>
<td>49%</td>
<td>NSDUH: 45%</td>
</tr>
</tbody>
</table>

NHIS: National Health Interview Survey; NSDUH: National Survey on Drug Use and Health

Table 2: High missing rates of All of Us survey questions

<table>
<thead>
<tr>
<th>Question</th>
<th>Missing rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you have completely stopped smoking cigarettes, how old were you when you stopped?</td>
<td>13%</td>
</tr>
<tr>
<td>If Yes to hysterectomy, age at surgery?</td>
<td>21%</td>
</tr>
<tr>
<td>If Yes to ovary removal, age at surgery?</td>
<td>29%</td>
</tr>
</tbody>
</table>

While the survey measures generally performed as expected with no major causes for concern, we identified a few data quality concerns affecting small amounts of the data. These included unanticipated branching logic behavior; failed field validation; and duplicate data transmissions. Each issue identified was meticulously investigated, documented, and communicated to the appropriate data management teams for resolution. Resolutions ranged from introducing additional validation checks into the survey deployment process to addressing technical configuration errors and amending the affected data.

Conclusion

This project is the first look at health survey data from the All of Us Research Program. This analysis helped confirm the integrity of health survey data for the initial three surveys of the program. We learned that the surveys are collecting data largely in line with expectations and have been technically implemented in accordance with intended survey operations. We also detected a small number of issues that offer an opportunity for remediation and improvement in data quality, including missingness in questions that targeted temporality. Compared to previous studies from which questions were derived, the survey questions are being answered by participants largely in an expected manner and, at this point, there were minimal changes recommended to some of the measures with high missingness. These changes included adding the word “about” to the age questions and adding text prior to certain questions to explain the importance of answering these questions. We will continue to monitor the performance of All of Us surveys over time, with particular interest in examining the utility of newly release surveys and exploring changes in response frequencies as the cohort evolves. By analyzing existing data from health surveys, All of Us can refine and adapt questions to improve their completeness, clarity, and provide accurate, cleaner data that researchers can use.

There were some limitations while conducting analysis, revealing areas for improvement and future directions. First, the population was skewed towards individuals identifying as white, female and/or with higher education attainment. This skewed population may have led to the differences in frequencies shown in Table 1. Analysis of a larger, more diverse population will likely generate different response distributions. Second, free text field responses were not included in the analysis due to the risk of receiving personally identifiable information. In the next iteration of assessment, free text fields will be analyzed to determine if other options should be added to existing response choices. Third, these analyses did not explore relationships between different characteristics, such as smoking relative to age or gender. Future analyses could look more closely at potentially dependent characteristics.

References

Addressing Diagnostic Errors Proactively using Electronic Events to Mitigate Harm during Inpatient Episodes of Care

Anuj K Dalal, MD,1,2 Kumiko Schnock, RN PhD,1 Nick Piniella,1 Kerrin Bersani,1 Pamela Garabedian, MS,1 Kevin Carr, MS, Ronen Rozenblum, PhD, MPH,1,2 Stuart Lipsitz, PhD,1,2 Jacqueline A Griffin PhD,3 Jeffrey L Schnipper, MD, MPH,1,2 David W Bates, MD, MS1,2

1Brigham and Women’s Hospital, Boston, MA, 2Harvard Medical School, Boston, MA 3Northeastern University, Boston, MA

Introduction: Diagnostic error in acute care represents an unresolved safety issue: error rates range from 4.8 to 49.8%. If the diagnosis is delayed or incorrect, the patient may not get correct treatment in a timely manner. Underlying contributing factors include system flaws (e.g., communication barriers) and cognitive errors (e.g., anchoring), factors that are often overlooked by workflow and technological solutions. Recent work (1) describing events typically captured by the electronic health record (EHR) has yet to be operationalized as “e-triggers” to identify cases of actual diagnostic error, as well as predictors that proactively identify patients at risk for preventable harm due to diagnostic error. As part of our new AHRQ-funded Patient Safety Learning Laboratory, a collaboration between Brigham and Women’s Hospital and Northeastern University, we aim to address diagnostic error in hospitalized patients by using these electronic events.

Methods: The study is being conducted at a large academic medical center in Boston, MA. Our team, comprised of experts in systems engineering, human factors, patient safety, quality improvement, clinical informatics, hospital medicine, and advanced statistics, is employing a sociotechnical approach to guide each project phase: problem analysis, design and development, and implementation and evaluation. Specifically, we are using the Safer Dx framework (2) to identify pertinent research questions across sociotechnical work system and diagnostic process dimensions, and to identify key activities and workstreams during each project phase.

Results: During the problem analysis phase, we identified three core activities (Table 1) in which we utilized systems engineering and human factors methods to analyze cases, understand stakeholders’ perspectives of actual diagnostic error, and validate previously described electronic events. First, we conducted a representative case analysis in which we identified cases of actual diagnostic error for 14 conditions (e.g., pneumonia) and undifferentiated symptoms (e.g., abdominal pain) from our hospital’s quality assurance committee, morbidity and mortality case reviews, safety reporting system, and patient and family relations department. We used Ishikawa and fault-tree analyses to understand system and cognitive factors that led to harm in these cases, and then prioritize key diagnostic process failures. Second, we conducted a qualitative analysis in which we performed structured interviews with key stakeholders and patient advisors to understand their perceptions of diagnostic error. Third, we conducted a quantitative analysis in which we retrospectively identified cases of potential diagnostic error by applying various electronic screening criteria to previously hospitalized patients. Currently, we are determining the sensitivity, specificity, and positive predictive value of these screening criteria by confirming the presence or absence of diagnostic error via chart review and a clinician adjudication process.

Discussion: We are using the knowledge generated from these core analytic activities to design and develop a robust digital health intervention (Figure 1, left) that will serve to engage patients and clinicians to ensure that clinical trajectories match the anticipated course for working diagnoses or symptoms, and align with patient and clinician expectations. We are using a user-centered approach to identify requirements according to key diagnostic process dimensions (Figure 1, right) and to improve utilization of existing EHR functionality (e.g., problem lists). We are also determining how specific electronic events can be used to identify potential cases of diagnostic error which will enable our team to evaluate the impact of the intervention on an enriched cohort of patients with actual diagnostic error. Our experience to date underscores the complexity and interdisciplinary expertise required to design strategies that leverage patient-reported and EHR data in real-time to enable front-line clinicians to proactively identify patients at risk for diagnostic error. During implementation and evaluation phases, our goal is to demonstrate a reduction in the rate of diagnostic error for patients admitted to general medicine units.

Acknowledgements: The PSLL project is supported by funding from AHRQ (R18-HS026613).

Table 1. Core Activities, Goals, and Methods used to Analyze Diagnostic Error in Acute Care

<table>
<thead>
<tr>
<th>Core Activity</th>
<th>Goals &amp; Methods</th>
</tr>
</thead>
</table>
| **Representative Case Analysis** | • Using medical record review and critical incident interviewing to identify representative cases  
• Using Ishikawa diagrams to identify key system and cognitive factors that lead to potential harm for 14 representative symptoms, clinical states, and conditions: chest pain, abdominal pain, altered mental status, dyspnea, failure to thrive, cough, anticoagulation, venous thromboembolism, asthma and chronic obstructive pulmonary disease, pneumonia, protein-calorie malnutrition, sepsis, cellulitis, soft tissue infection  
• Using fault tree analysis to understand and prioritize key failures in the diagnostic process. |
| **Qualitative Analysis**   | • Conducting semi-structured interviews to understand clinicians’ perspectives of diagnostic error: 1) general perceptions of diagnostic error; 2) types of system and cognitive factors associated with diagnostic error; 3) clinician behaviors associated with diagnostic error: lack of attentiveness to patient symptoms, questions, and concerns; disrespecting patients; failing to communicate; manipulation or deception  
• Conducting task analysis and workflow observations to identify key activities related to each diagnostic process dimension: admission/initial encounter, diagnostic test performance/interpretation, follow-up and tracking of diagnostic information, subspecialty consultation, patient experience, care team communication/collaboration |
| **Quantitative Analysis**  | • Validating screening criteria to identify cases of diagnostic error: transfers to ICU; discrepancies in observed vs expected LOS; unexpected surgery or procedure; code or rapid response; change of code status or diagnosis code; multiple medication changes or consultations; abrupt stop in medication; cancelled procedure  
• Using statistical methods to calculate, sensitivity, specificity, and positive predictive value for each criterion  
• Using risk assessment and prioritization to determine criteria that have greatest impact once operationalized |

Figure 1. Intervention Design & Development: Potential Use of EHR-Integrated Digital Health Tools to Address Key Diagnostic Process Failures

We are using systems engineering and human factors approaches to design, develop, and implement a suite of EHR-integrated digital health tools (left) to address high frequency errors in the diagnostic process (right) identified during the problem analysis phase. **Patient-facing health IT tools** (A, B) empower patients (or care partners) to visualize, understand, and self-report concerns about their care. **Provider-facing health IT tools** (A, C, D) reliably warn clinicians of potential threats to safety in real-time from both the patient’s and clinician’s perspective. Potential solutions will leverage both patient-facing (web-based survey tools to empower real-time self-reporting) and clinician-facing (dashboards, patient-centered communication apps, smart notifications) to address the system and cognitive failures in the diagnostic process that may lead to error. For example, at admission, our quality and safety dashboard could use EHR data services to identify newly admitted patients in real-time who have potential risk factors for diagnostic error (e.g., an unplanned primary care or ED visit or multiple ED/urgent care visits within 2 weeks prior to hospitalization, etc.). To evaluate the impact of intervention on the rate of diagnostic error, we are also developing a recent retrospective review process to identify actual cases of diagnostic error (i.e., an unplanned primary care or ED visit or multiple ED/urgent care visits within 2 weeks prior to hospitalization, etc.). To evaluate the impact of intervention on the rate of diagnostic error, we are also developing a recent retrospective review process to identify actual cases of diagnostic error (i.e., an unplanned primary care or ED visit or multiple ED/urgent care visits within 2 weeks prior to hospitalization, etc.). To evaluate the impact of intervention on the rate of diagnostic error, we are also developing a recent retrospective review process to identify actual cases of diagnostic error (i.e., an unplanned primary care or ED visit or multiple ED/urgent care visits within 2 weeks prior to hospitalization, etc.). To evaluate the impact of intervention on the rate of diagnostic error, we are also developing a recent retrospective review process to identify actual cases of diagnostic error (i.e., an unplanned primary care or ED visit or multiple ED/urgent care visits within 2 weeks prior to hospitalization, etc.).
A frame semantic overview of NLP-based information extraction for cancer-related EHR notes: a scoping review

Surabhi Datta, MS¹, Elmer V. Bernstam, MD, MSE¹,², Kirk Roberts, PhD¹
¹School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
²Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, TX

Introduction

Unstructured clinical data about cancer patients is increasingly available in clinical notes in Electronic Health Records (EHRs) and other systems. There is increasing interest in utilizing this data for biomedical research. Many publications argue for the need to capture important clinical information on cancer within the free text information in EHRs, ranging from information about tissue specimens¹ to disease-related and outcome information², thus facilitating translational research by associating molecular information with disease phenotype. Some works are more focused on extracting information to improve cancer screening efficiency³. All these works have focused on deriving cancer-related information automatically using various natural language processing (NLP) techniques. More recently, deep learning-based NLP systems are gaining prominence for cancer information extraction from EHRs⁴. However, many of these researchers put sizable effort into designing and implementing NLP systems that extract similar information types. In this scoping review, we investigate the cancer information types and organize the extracted information into frames, based on the linguistic theory of ‘frame semantics’⁵. We review existing literature for clinical NLP related to cancer, and demonstrate why a frame semantic perspective is important. Our aim is to provide a list of cancer-related frames in existing work that would be valuable to the scientific community.

Methods

We obtained a total of 899 articles using keywords such as ‘natural language processing’ or ‘NLP’ anywhere in the article, and one of the keywords from ‘cancer’, ‘tumor’, and ‘oncology’ in the title from PubMed, Google Scholar, ACL Anthology, and two existing reviews. Results were limited to papers published between January 1, 2000 and September 20, 2018. We included articles if the title or abstract contained the description of an NLP method to extract cancer-related information from EHR notes. This resulted in selecting 173 papers, after which two further exclusion criteria were applied to the full text. We excluded articles where the extraction techniques used were too broad to be represented as frames (e.g., document classification) and also where very low-level extraction methods were used (e.g. identifying all clinical concepts in a cancer note). Finally, 78 articles were included in the review. We organized this information according to frame semantic principles to help identify common areas of overlap and potential gaps.

Results

Frames were created from the reviewed articles pertaining to cancer information such as cancer diagnosis, tumor description, cancer procedure, breast cancer diagnosis, prostate cancer diagnosis, and pain in prostate cancer patients. These frames included both a definition as well as specific frame elements (i.e. extractable attributes). We also constructed relationships between frames (shown in Figure 1). The different colors in Figure 1 correspond to the various broader categories of information type. We defined three kinds of frame relations based on those found in Berkeley FrameNet⁶. We found that cancer diagnosis was the most common frame among the reviewed papers (36 out of 78), with recent work focusing on extracting information related to treatment and breast cancer diagnosis. We also noted the distribution of papers at the frame element level, which reflects the important clinical attributes related to cancer that interest NLP researchers. The five most extracted frame elements along with the number of referenced articles are presented in Table 1.

Discussion

Our scoping review provides a detailed overview of the current research in the cancer information extraction domain from unstructured EHR notes using NLP. We conducted the review from a frame semantic perspective, described various frames along with their elements as well as examined the relations between frames. Since many researchers are trying to extract similar frames or frame elements (though not always using the language of frame semantics), this review can help develop a general-purpose cancer frame resource and NLP system that would extract a broad range of important cancer information types.
Acknowledgements

This work was supported by the U.S. National Library of Medicine under award R00LM012104, and the Cancer Prevention Research Institute of Texas (CPRIT) under award RP170668. The authors would like to thank Funda Meric-Bernstam for reviewing the work and suggesting clarifications from an oncology perspective.

References

Collecting Individual-Level Social Determinants of Health to Inform Patient-Centered Outcomes Research in Mental Health

Joseph J Deferio, MPH1; Annie C Myers, MA1; John M. Meddar1, Judith Cukor, PhD2, Jyotishman Pathak, PhD1

1Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY; 2Department of Psychiatry, Weill Cornell Medicine, New York, NY

Introduction: According to a 2018 World Health Organization (WHO) report, depression is listed as the largest global contributor to disability and is also a known contributor to suicide deaths.1 Global efforts aimed at mitigating depression prevalence recognize the importance of a multimodal approach to combating poor mental health. Recent studies suggest that providing adequate access to mental healthcare alone will not attenuate depression prevalence rates, but rather through addressing social determinants of health (SDOH), more pronounced and sustainable changes can be attained.2 Despite increased interest from health systems, public health agencies, and community organizations in understanding and addressing people’s social, economic, and environmental circumstances, individual-level SDOH collection efforts are uncommon, typically ad-hoc and disjointed, and there remains no firm understanding about which SDOH elements patients are even willing to share with their healthcare provider, health system, or publicly. In response, a number of health systems have developed tools to facilitate individual-level SDOH data collection, but their implementation is not yet widespread and they do not collect increasingly relevant online socio-behavioral estimates. The purpose of the study is to elicit from patients, in a structured and easily-deployable manner, information about experiences with depression/mental illness and how one’s particular social (physical and digital) and environmental conditions may or may not influence their mental health. We aim to use these results to design future observational and prospective studies that link individual-level SDOH data to other data sources, such as electronic health records (EHRs).

Methods: We designed a brief, anonymous, online survey to understand patient attitudes about sharing mental healthcare utilization, behavior, treatment, online/social media use, and general social determinants (factors) of health (http://bit.ly/wcmc_sdh). Our survey was constructed using 9 questions from existing instruments, such as the Protocol for Responding to and Assessing Patients’ Assets, Risks, and Experiences (PRAPARE) tool, designed to measure SDOH as well as structural and attitudinal barriers to healthcare.3, 4 From the existing PRAPARE tool, we drew questions that focused on structural environment such as housing situation, number of people living in household, and whether they felt physically and emotionally safe in their neighborhood. In addition, we created 28 new questions to describe the patient’s perception of neighborhood characteristics such as, but not limited to, safety, green space, access to public transportation, and blight. The survey is targeted towards patients that are currently experiencing depression or have a history of major depressive disorder (MDD) (including bipolar disorder), as defined by diagnosis codes or a PHQ-9 score ≥ 10. The survey is comprised of a total of 45 questions (multiple choice and free text), and can be completed in-person on a tablet computer set up specifically for the study at the health system clinic during their appointment, or at the patient’s leisure via a shareable hyperlink. All responses are automatically uploaded to a secure, HIPAA compliant database. This study has been approved by the Weill Cornell Medicine Institutional Review Board (#171008654).

Table 1 Participant Demographics

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Participants (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19 (41.3)</td>
</tr>
<tr>
<td>Female</td>
<td>27 (58.7)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td>18-34</td>
<td>27 (58.7)</td>
</tr>
<tr>
<td>35-54</td>
<td>10 (21.7)</td>
</tr>
<tr>
<td>55+</td>
<td>9 (19.6)</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>3 (6.0)</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>9 (18.0)</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>29 (58.0)</td>
</tr>
<tr>
<td>Black/African-American</td>
<td>7 (14.0)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (4.0)</td>
</tr>
<tr>
<td><strong>Income</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;$25,999</td>
<td>5 (10.9)</td>
</tr>
<tr>
<td>$26,000-$51,999</td>
<td>11 (23.9)</td>
</tr>
<tr>
<td>$52,000-$74,999</td>
<td>7 (15.2)</td>
</tr>
<tr>
<td>$75,000-$99,999</td>
<td>6 (13.0)</td>
</tr>
<tr>
<td>$100,000+</td>
<td>12 (26.1)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td>9 (19.0)</td>
</tr>
<tr>
<td>Bachelor’s</td>
<td>27 (58.7)</td>
</tr>
<tr>
<td>Master’s</td>
<td>6 (13.0)</td>
</tr>
<tr>
<td>Doctorate</td>
<td>3 (6.5)</td>
</tr>
<tr>
<td><strong>Insurance Status</strong></td>
<td></td>
</tr>
<tr>
<td>Medicare/Medicaid</td>
<td>4 (8.7)</td>
</tr>
<tr>
<td>Employer coverage</td>
<td>31 (67.4)</td>
</tr>
<tr>
<td>Self-purchased</td>
<td>11 (23.9)</td>
</tr>
</tbody>
</table>
**Results:** As of July 2019, we have collected data from 47 adult patients out of an initial recruitment target of 200 (see Table 1 for study population characteristics). On average, it took 9 minutes for patients to complete the survey. The survey respondents thus far include 19 men and 27 women (1 participant declined to report gender). More than half of participants (58.7%) reported being between the ages of 18 and 34.

89.4% of participants reported that daily events at home or in their community have caused increased sadness in the past 12 months. Nearly all participants (87%) have felt emotionally isolated from friends, family, and their community at some point in the past 12 months, with 19.6% feeling emotionally isolated nearly every day. When asked to share specific social and environmental triggers, common themes included pressure from relationships, work, and finances.

Our preliminary findings show that when seeking mental healthcare, 36.2% of participants reported facing difficulty and barriers to access. The most prominent barriers to mental healthcare were structural and included the uncertainty of where to go or which healthcare provider to see, and the financial burden of treatment. 72.3% of participants reported that they looked online or at social media for guidance before seeking care, with one-third relying on non-academic online resources such as WebMD. 85.3% reported looking online for three months to more than one year before deciding to seek professional help.

Regarding neighborhood safety, 65.2% of participants reported that they feel comfortable walking alone in their neighborhood during the day and night, while 30.4% reported only feeling comfortable during the day time. More than half (67.4%) reported feeling physically and emotionally safe in their current living situation, although 8.7% participants reported they do not feel safe, and 21.7% reported feeling unsure about their safety in their neighborhood. The majority (80.4%) felt that having access to public transportation has made them feel more connected to their city and community, and enabled improved socialization.

**Discussion:** This study was initiated because there remains a gap in knowledge about individual-level social determinants of mental health. While results are still preliminary, we have shown that patients are willing to share intimate details about different aspects of their lives, such as how they have dealt with their mental illness, social difficulties, and perceived neighborhood safety. We have also seen that relevant SDOH information can be collected from patients in a short amount of time, and at the patient’s convenience, thus requiring less of the healthcare provider’s time. Although we are currently focusing on individuals with MDD, findings from this study can also provide insight into the social factors that may be affecting the mental health population as a whole. Future integration with EHRs and other external datasets is another goal of this work. Survey tools like PRAPARE provide EHR templates which facilitate integration into patient clinical records, therefore it is possible for other such tools to do the same. Our study, as of July 2019, is limited by the number of participants at a single institution. In addition, the study sample suffers from bias in the population makeup, including higher SES and education, commercially insured, and those who are already receiving mental health care. Lastly, the study does not address issues on whether or not patients would consent to the collection of this data if the survey also included protected health information (PHI), and their results were linked to the EHRs.

**Acknowledgments:** This research has been supported in part by funding from NIH R01 MH105384.

**References**


Comparison of Free-Text Synthetic Data Produced by Three Generative Adversarial Networks for Collaborative Health Data Analytics

Gregory P. Dexter¹, Shaun J. Grannis MD, MS²,³ Suranga N. Kasthuriratne, PhD²,³
¹Purdue University, West Lafayette, IN, USA; ²Regenstrief Institute, Indianapolis, IN, USA; ³Indiana University, Indianapolis, IN, USA

Introduction. Fragmentation of Health Information Systems (HIS) and legal restrictions on sharing Patient Health Identifiers (PHI) limits access to healthcare data for research purposes. This impedes (a) sharing and re-use of Machine Learning (ML) solutions across larger audiences, (b) inter-organizational collaboration to address various healthcare challenges, and (c) building of generalized ML solutions targeting large geographic areas or multiple health systems. The advent of Generative Adversarial Networks (GAN)¹ capable of mimicking real-valued data present considerable potential to democratize access to health data. These approaches have already been used to generate numerical clinical data that is comparable to real data². However, valuable healthcare information is also trapped within free-text datasets. We leverage enhanced GAN algorithms that support free-text to generate high-quality synthetic free-text healthcare datasets capable of being used to train decision models that perform similarly to models trained using real data, and present limited re-identification risk.

Materials and methods. We extracted 6,770 salmonella laboratory test reports (17.91% positive, 82.08% negative) from the Indiana Network for Patient Care (INPC), a large statewide Health Information Exchange (HIE). The reports were split into train (90%) and holdout (10%) report sets. We used the training report sets to train SeqGAN, MaliGAN and RankGAN adversarial models³ using hyperparameter tuning techniques. Next, we generated synthetic datasets from each optimally trained adversarial model. We evaluate the quality of synthetic datasets by comparing them to real data using Bilingual Evaluation Understudy (BLEU) 1-4 and Google-BLEU scores⁴ which measure n-gram overlap between datasets. To evaluate their ability to replicate ML efforts, we vectorized the three synthetic report sets and the real (training) report set using methods described in a previous study⁵. We extracted the top 5, 10, 15, and 20 features from each of the four vector sets using Gini impurity. For each data vector and feature subset size, we built Random Forest⁶ decision models predicting a positive or negative outcome for salmonella, and assessed its performance using the holdout data vector. Re-identification risk was evaluated using techniques described by Choi et. al.² to calculate presence disclosure, which evaluates an attacker’s ability to determine if any given report was used to train a GAN model. Figure 1 presents our research process.

Results. BLEU 1-4 and Google-BLEU scores comparing real and synthetic data across positive and negative report sets demonstrate that SeqGAN and MaliGAN models were superior to RankGAN generated synthetic data (table 1). Analysis of the top 5, 10, 15, and 20 features extracted from each of the three synthetic report sets and the real report set determined that correlation between RankGAN vs. real features (only 30%-60%) was inferior to SeqGAN and MaliGAN vs. real features (~60% correlation) (table 2). When tested using the holdout dataset, models trained using MaliGAN and RankGAN generated data underperformed in comparison to models trained using real data. However, models trained using 10 or more features derived from SeqGAN generated synthetic data achieved performance measures that were statistically similar to measures reported by models trained using real data (figure 2). Further, presence disclosure assessment of SeqGAN based synthetic data using various hamming score thresholds determined limited re-identification risk.

Discussion. Our results inform two challenges; use of GAN models to generate synthetic unstructured free-text healthcare data with limited re-identification risk, and use of this data to enable collaborative research and re-use of decision models and analytical solutions. We found that SeqGAN generated synthetic data was superior to those generated by other GAN models, and suitable for creating synthetic free-text data that can be used to replicate ML solutions. Thus, an organization that possesses rich free-text data sources, but lacks adequate ML expertise can leverage our approach to generate synthetic data with low re-identification risk. This data could be shared with external experts who would use it to train machine learning models. Once optimal models have been identified, they could be applied to the original dataset with compatible performance measures. However, our study did not include any analysis of the readability or syntactic/grammatical accuracy of the synthetic reports. As such, these synthetic datasets are only suitable for ML purposes, and not as teaching or learning resources. Next steps include (a) manual assessment of the readability and syntactic correctness of synthetic reports using human experts (Turing test), and (b) investigation of other word and grammar-based measures that inform assessment of synthetic data.
Figure 1. A workflow depicting our study approach from laboratory report extraction to decision model evaluation.

Table 1. BLEU 1-4 scores comparing real vs. synthetic reports

<table>
<thead>
<tr>
<th>Measure</th>
<th>Positive reports (synthetic vs. real)</th>
<th>Negative reports (synthetic vs. real)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqGAN</td>
<td>BLEU-1 0.913</td>
<td>0.944</td>
</tr>
<tr>
<td></td>
<td>BLEU-2 0.675</td>
<td>0.742</td>
</tr>
<tr>
<td></td>
<td>BLEU-3 0.480</td>
<td>0.552</td>
</tr>
<tr>
<td></td>
<td>BLEU-4 0.331</td>
<td>0.409</td>
</tr>
<tr>
<td>MaliGAN</td>
<td>BLEU-1 0.955</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>BLEU-2 0.472</td>
<td>0.772</td>
</tr>
<tr>
<td></td>
<td>BLEU-3 0.251</td>
<td>0.603</td>
</tr>
<tr>
<td></td>
<td>BLEU-4 0.133</td>
<td>0.458</td>
</tr>
<tr>
<td>RankGAN</td>
<td>BLEU-1 0.841</td>
<td>0.905</td>
</tr>
<tr>
<td></td>
<td>BLEU-2 0.313</td>
<td>0.657</td>
</tr>
<tr>
<td></td>
<td>BLEU-3 0.124</td>
<td>0.468</td>
</tr>
<tr>
<td></td>
<td>BLEU-4 0.061</td>
<td>0.334</td>
</tr>
</tbody>
</table>

Figure 2. Performance measures for decision models trained using real and SeqGAN generated data (feature size = 5, 10, 15, 20)

Table 2. List of top 20 feature stems extracted from the real and SeqGAN report sets (bold tokens appear in both feature sets).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>List of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>shigella, isol, salmonella, speci, state, indiana, depart, chslb, sct, group, azithromycin, sent, enter, stool, typhi, coli, preliminari, klebsiella, absent, pathogen</td>
</tr>
<tr>
<td>SeqGAN</td>
<td>shigella, salmonella, speci, isol, campylobact, health, indiana, suscept, group, typhi, enter, stool, depart, final, ml, confirm, coli, non, chslb, sent</td>
</tr>
</tbody>
</table>

References

Local and State Public Health Informatics Workforce Skills and Needs: A Descriptive Analysis using the PH WINS

Brian E. Dixon1,2, Timothy D. McFarlane1, Shaun J. Grannis2,3 P. Joseph Gibson1,4
1Indiana University Fairbanks School of Public Health, Department of Epidemiology, Indianapolis, IN; 2Regenstrief Institute, Center for Biomedical Informatics, Indianapolis, IN; 3Indiana University School of Medicine, Indianapolis, IN; 4Marion County Public Health Department, Indianapolis, IN

Introduction

In 2014, the Association of State and Territorial Health Officials and the de Beaumont Foundation fielded the Public Health Workforce Interests and Needs Survey (PH WINS), the first nationally representative sample of state agency (SHA) public health workers in the United States. Using data from that survey, Dixon and colleagues (1) estimated that public health informatics (PHI) specialists constitute 1.3% of the SHA workforce.

While the proportion of public health workers identifying as PHI specialists is small, informatics needs agency-wide are large for health departments to become “informatics-savvy.” As a result, informatics skills have been added to essential competencies across public health disciplines (2), and the Council on Education for Public Health has included informatics as a cross-cutting competency for master of public health programs (3).

In 2017, the PH WINS was fielded a second time, and the survey included a representative set of local public health department (LHD) workers in addition to SHA workers. Using data from the newer survey, we sought to update our prior analysis of the PHI workforce as well as assess the informatics needs of the broader public health workforce.

Methods

We performed a cross-sectional study using the nationally representative 2017 PH WINS. Details on the PH WINS sampling and data set construction can be found in the article by Leider et al. (4). The complete analysis summarized in this abstract, approved by the Institutional Review Board at Indiana University, can be found in the article by McFarlane et al. published in the Journal of Public Health Practice & Management (5).

Respondents’ perceived skills and needs were captured using 21 questions across 8 “strategic skill” domains. We focused on questions addressing core public health competencies possessing the greatest overlap with existing PHI competencies (6), including: identifying appropriate sources of data and information to assess the health of a community; collecting valid data for use in decision making; participating in quality improvement processes for agency programs and services; and identifying evidence-based approaches to address public health issues.

Respondents were asked to rate selected public health competencies with respect to the importance to their day-to-day work (ie, not important, somewhat unimportant, somewhat important, and very important) and their current skill level (eg, not applicable, unable to perform, beginner, proficient, expert). To identify the highest priorities for future workforce training, we present skill gaps. A skill gap was defined as discordance between self-reported importance (ie, need) and skill level, for example, those reporting the competency as “somewhat important” or “very important” and “unable to perform” or “beginner.”

Results

Of all public health employees surveyed, informaticians accounted for 1.1% (0.9%-1.2%, 95% CI) of SHA respondents and 0.5% (0.2%-0.7%, 95% CI) of LHD respondents working in a Big City Health Coalition agency, those that serve the top 30 most populous urban areas in the United States.

Table 1 below details discordant skills gaps by role and setting for two key informatics competencies. Informaticians reported the lowest discordance in all categories (with one notable exception in LHD), and public health workers overall were more likely to report a high need, low skill discordance than the reverse. The high need, low skill discordance represents a shortage of key skills in the workforce.

Table 1. Weighted Percentages for Selected Core Public Health Competencies among Public Health Roles
Identify appropriate sources of data and information to assess the health of a community

Collect valid data for use in decision making

<table>
<thead>
<tr>
<th></th>
<th>Low Need, High Skill</th>
<th>High Need, Low Skill</th>
<th>Low Need, High Skill</th>
<th>High Need, Low Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>w% (se)</td>
<td>n</td>
<td>w% (se)</td>
<td>n</td>
</tr>
<tr>
<td>SHA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHI</td>
<td>2</td>
<td>1.4 (1.1)</td>
<td>12</td>
<td>8.3 (2.6)</td>
</tr>
<tr>
<td>IT</td>
<td>23</td>
<td>7.0 (1.7)</td>
<td>48</td>
<td>18.6 (4.0)</td>
</tr>
<tr>
<td>PHS</td>
<td>186</td>
<td>2.9 (0.3)</td>
<td>851</td>
<td>16.1 (0.6)</td>
</tr>
<tr>
<td>CL</td>
<td>48</td>
<td>2.5 (0.4)</td>
<td>433</td>
<td>22.9 (1.0)</td>
</tr>
<tr>
<td>BCHC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHI</td>
<td>0</td>
<td>0 (0)</td>
<td>4</td>
<td>17.8 (12.8)</td>
</tr>
<tr>
<td>IT</td>
<td>3</td>
<td>5.8 (3.1)</td>
<td>12</td>
<td>22.1 (7.7)</td>
</tr>
<tr>
<td>PHS</td>
<td>64</td>
<td>2.8 (0.3)</td>
<td>343</td>
<td>16.9 (1.0)</td>
</tr>
<tr>
<td>CL</td>
<td>38</td>
<td>2.7 (0.5)</td>
<td>320</td>
<td>22.5 (1.5)</td>
</tr>
<tr>
<td><strong>P-value</strong></td>
<td>0.0021</td>
<td>&lt;0.0001</td>
<td>NR</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

1 Rao-Scott chi-square for the distribution of skills gap; NR = not reported due to zero cell.

SHA = State Health Agency; BCHC = Big City Health Coalition

Public Health Roles Based on Self-Report: PHI = Public Health Informatics; IT/IS = Information Technology Specialist or Information Systems Manager; PHS = Public Health Science; CL = Clinical and Laboratory.

Discussion

An informatics-savvy health department requires PHI competencies not just among specialists (e.g., informaticians) but also among front line workers, program area managers, and executive leadership. While encouraging that informaticians generally reported having the skills needed for their jobs, our findings suggest that additional informatics training is necessary for the broader public health workforce. Discordance suggests that health departments should examine ways to enhance training for PHI-related competencies. AMIA and its members could work collaboratively with public health associations and schools to create opportunities for informatics competency training for existing workers in the public health workforce.

Acknowledgements

PH WINS was funded by the de Beaumont Foundation and conducted by the Association of State and Territorial Health Officials and the de Beaumont Foundation. ASTHO provided the PH WINS data and methodological support to the authors as part of an initiative to analyze and disseminate findings from the 2017 fielding of the survey. The sponsors did not have any role in the development of this abstract.

References

Towards a Universal Document-Level Clinical Text Encoder: Methods for Neural Network Pre-training with Applications to Substance Misuse

Dmitriy Dligach, PhD\textsuperscript{1}, Majid Afshar\textsuperscript{1}, MD, MSCR, Timothy Miller, PhD\textsuperscript{2}
\textsuperscript{1}Loyola University Chicago, Chicago, IL; \textsuperscript{2}Boston Children's Hospital and Harvard Medical School, Boston, MA

Introduction

Recent neural network models have shown state-of-the-art results on a number of natural language processing (NLP) benchmarks and even human-level performance on several narrowly-defined tasks such as question answering and machine translation. Yet, this success required tens or hundreds of thousands of labeled examples. Procuring annotated datasets of this size is not feasible for most tasks in clinical NLP due to the high cost of manual labeling.

This problem of insufficient training data is addressed in computer vision by means of pre-training classifiers on massive datasets such as ImageNet and subsequently refining them on a more specialized classification task. Recent years have seen a number of new approaches in NLP for word- and sentence-level pre-training. In this work, we investigate a form of pre-training that utilizes billing code prediction objectives; because billing codes are available in abundance in health care institutions and are linked to document-level entities, such as clinical encounters, we are able to move beyond sentence-level encoding to document level by means of supervised pre-training.

We investigate a simple text encoder that takes UMLS concepts or raw text as input and is trained using a billing code prediction objective. The encoder is subsequently used to generate patient representations that succinctly capture patient information. We also introduce a novel phenotype-specific encoder that makes it possible to trade the generality of the resulting text representations for better performance on a specific phenotyping task.

We evaluate our encoder on several phenotyping tasks using labeled datasets. First, in the interests of reproducibility, we evaluate on the publicly available i2b2 comorbidity challenge data, establishing a new state-of-the-art. We also apply our encoder to two novel and high-impact substance misuse tasks, predicting opioid and alcohol misuse in trauma patients.

Materials and Methods

We explore several neural architectures that work directly with text and simple named entity features automatically extracted from text. The first encoder is a deep averaging network (DAN) that takes a set of Unified Medical Language System (UMLS) concept identifiers (CUIs), maps them to their 300-dimensional embeddings, averages them, and projects them to the penultimate fully-connected hidden layer, encoding the input as a fixed-sized dense vector. During pre-training, the final (output) network layer consists of $n$ sigmoid units, each representing a unique billing code. The second encoder is a convolutional neural network (CNN) that operates directly on the text of the notes. The embedding layer is followed by a convolutional layer, a max pooling layer, and a fully-connected layer. The output layer is identical to the DAN architecture above.

Both encoders are trained using binary cross-entropy loss function and RMSProp optimizer to jointly predict billing codes. To use the patient encoder as a feature extractor, we freeze the network weights, push the text of the notes through the network, and collect the computed values of the hidden layer nodes, thus obtaining a dense vector representing the input text, that can be used as input for any machine learning task, e.g. to train a supervised classifier.

While our ultimate goal is learning a universal clinical text encoder, we also observe that it is possible to train a phenotype-specific encoder by restricting the billing code prediction targets to a set that is relevant to a specific phenotype. While a universal encoder could be more practical, since it needs to be pre-trained only once, a phenotype-specific encoder could potentially perform better for a specific medical condition. We train two phenotype-specific text encoders for detecting substance misuse status, by modifying the encoder's training objective to predict only the codes associated with two substance misuse scenarios: alcohol and opioid misuse.

Experiments

We pre-train all text encoders using the MIMIC III \cite{6869554} corpus. MIMIC III contains notes and structured data for over 40,000 Beth Israel Deaconess Medical Center critical care patients. To evaluate the quality of an encoder, we deploy it as a feature extractor to generate text representations we can use as input to a linear SVM classifier. To obtain a vector representing a patient, we freeze the network weights and push the patient text through the encoder, harvesting...
the computed values of the units of one of the intermediate network layers. We compare all models to a baseline SVM classifier that we train for each phenotype with bag-of-cui features.

For evaluation, we use a publicly available dataset from the i2b2 Obesity challenge [2], which consists of 1237 discharge summaries from the Partners HealthCare annotated with respect to obesity and its fifteen most common comorbidities. In addition, we use two in-house substance (alcohol and opioid) misuse datasets developed at the Loyola University Medical Center. We emphasize that the patient data we use in our evaluation originates from healthcare institutions (Partners HealthCare and Loyola) that are different from the ones on which the encoders were trained (Beth Israel). This evaluation is challenging yet it presents a true test of robustness of the proposed methods.

Results

Evaluation results are shown in Table 1. Our CNN-based clinical text encoder outperformed the bag-of-cuis baseline by a wide margin. Our DAN-based encoder outperformed the bag-of-cuis baseline and our previous encoder by a wide margin.

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Encoder input</th>
<th>SVM input</th>
<th>Macro P</th>
<th>Macro R</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>CUIs</td>
<td>bag-of-cuis</td>
<td>0.733</td>
<td>0.65</td>
<td>0.675</td>
</tr>
<tr>
<td>CNN</td>
<td>Words</td>
<td>CNN hidden layer</td>
<td>0.719</td>
<td>0.723</td>
<td>0.718</td>
</tr>
<tr>
<td>DAN</td>
<td>CUIs</td>
<td>DAN hidden layer</td>
<td>0.752</td>
<td>0.751</td>
<td>0.746</td>
</tr>
</tbody>
</table>

Table 1. Average SVM classifier performance on sixteen i2b2 comorbidity challenge phenotyping tasks. Performance is compared to SVM trained on a bag-of-cuis representation of input notes (baseline) vs. the representations derived from encoders pre-trained on billing code prediction tasks.

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Pre-training targets</th>
<th>SVM input</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>none</td>
<td>bag-of-cuis</td>
<td>0.838</td>
</tr>
<tr>
<td>DAN</td>
<td>all billing codes</td>
<td>DAN hidden layer</td>
<td>0.889</td>
</tr>
<tr>
<td>DAN</td>
<td>opioid-specific billing codes</td>
<td>DAN hidden layer</td>
<td>0.951</td>
</tr>
</tbody>
</table>

Table 2. Comparison of different input representations on the performance of an SVM classifier on the opioid misuse data. Bag-of-cuis input (baseline) is compared to the performance of the input obtained from a deep averaging network encoder pre-trained on different billing code prediction tasks.

Our DAN-based encoder showed the best performance on the comorbidity data and we proceeded by evaluating it on the opioid misuse data, where it helped to improve the classifier performance by over five points. Finally, when the encoder was pre-trained on the opioid-specific billing codes only, we obtained further improvements, outperforming the bag-of-cuis baseline by over eleven points. Similarly to the opioid misuse task, we find that the use of our text encoder helps to improve the classifier performance on the alcohol misuse detection task, although the size of the improvements is more modest (results not shown due to space constraints).

Conclusion

In general, we find that pre-training using billing codes is a viable route for pre-training. The representations generated by jointly predicting the billing codes associated with a patient encounter have properties of universal patient representations as they were beneficial for all the phenotyping tasks reported here. While phenotype-specific pre-training is beneficial, it is less practical since it requires additional effort tuning the encoder to a specific set of billing codes. Nevertheless, we find that this is a viable route for trading the generality of the pre-trained encoder for better performance on a specific phenotyping task.

References

Development of an Automated Provider-Specific “Learning List” within a Commercial Electronic Health Record

Robert El-Kareh, MD MPH MS1, Aaron Kemp, MS2, Samantha Hurst, PhD3

1Department of Biomedical Informatics, UC San Diego Health, La Jolla, CA; 2Information Services, UC San Diego Health, La Jolla, CA; 3Department of Family Medicine and Public Health, UC San Diego, La Jolla, CA

Introduction

Accurate and timely diagnoses and appropriate management plans are keys to avoiding patient harms from delays in effective therapy, exposure to ineffective treatments and unnecessary tests. Clinicians establish their skills during training and then we assume that clinical experience will enable them to improve their clinical skills and reduce errors over time. However, for this calibration to occur, practicing clinicians must get ongoing feedback of the outcomes of their decisions. In many hospital settings, our increasingly fragmented care models lead to multiple handoffs of care and significantly hinder this feedback. We sought to develop a system integrated into our electronic health record (EHR) to provide consistent feedback of patient outcomes through an automated and individualized “Learning List.”

Methods

We are taking a multi-level approach to understand the various perspectives of stakeholders involved in the “Learning List” that highlights important information or events across transitions of care. We conducted interviews with frontline clinicians and with patient safety leaders within our organization. Our participants to date have included four internal medicine (IM) attendings, two IM residents, three emergency medicine (EM) attendings, two EM residents, and three of our organization’s patient safety leaders.

Our early interviews centered on current practices and challenges related to following up on prior patients. Building on what we learned, iterative changes were made to expand later interviews to include participant opinions regarding: (1) the data elements that would be most helpful to include to remind clinicians about the context of their care with each patient; (2) the post-handoff events which would be most important to highlight; and (3) how the information would best be presented.

We implemented our prototype Learning List system using Crystal Reports (SAP SE, Walldorf, Germany) connected to the relational database for our Epic EHR (Epic, Verona, WI). The system identifies transitions of care using the Treatment Team construct within the EHR. In addition, through Structured Query Language (SQL) queries, the system identifies a set of post-handoff events that were previously identified to be potentially indicative of a diagnostic delay or error. This set includes events such as death, code blue, escalation of care, re-presentation to the emergency department, early transfer between medical and surgical services and others. The system is available to all providers on eligible service teams upon electronic request and is generated daily. Each provider’s Learning List contains all patients handed off by that provider during a set look-back timeframe.

Results

During our early interviews with IM and EM physicians, we obtained results similar to a prior study that surveyed residents about their post-handoff follow-up practices at our institution. We found that physicians had substantially varied practices to keep track of patients they thought warranted follow-up. Many of these practices required manual addition of specific patients to a list and were felt by subjects to be variably effective.

In our later interviews with physicians, we identified the data elements of highest value to our subjects to remind them of the patient encounters. These elements included patient name, age, gender, chief complaint, and date/time that the physician was on the treatment team. The information that subjects felt was valuable to display about the post-handoff course included the current principal problem for the patient along with the set of triggered events. Subjects felt that a graphical display using icons was helpful to quickly identify which patients experienced important events post-handoff. Based on this input, we updated our system to its current form (Figure 1).

Our interviews with local patient safety leaders provided additional organization-wide perspectives on the use of the Learning List. These interviews have included considerations of sharing learning opportunities across providers and expanding the system to support a wider range of provider types, such as medical students. In addition, the patient
safety leaders discussed the relationship between this system and a broader effort to continue to advance a culture of safety across the organization.

<table>
<thead>
<tr>
<th>Who is the patient?</th>
<th>What happened after handoff?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart</td>
<td>Sticky Note</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td>Jane A. Doe</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td>John B. Doe</td>
</tr>
<tr>
<td><img src="image4.png" alt="Image" /></td>
<td>John C. Doe</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td>Jane B. Doe</td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td>John D. Doe</td>
</tr>
<tr>
<td><img src="image7.png" alt="Image" /></td>
<td>Jane C. Doe</td>
</tr>
<tr>
<td><img src="image8.png" alt="Image" /></td>
<td>Jane D. Doe</td>
</tr>
</tbody>
</table>

Figure 1. Format of Automated Provider-Specific Learning List.

Discussion

The EHR serves several purposes and we strongly believe that one priority should be to facilitate ongoing provider education and calibration of decision-making. We describe the development of an automated and provider-specific “Learning List” system implemented at an academic medical center and integrated into a commercial EHR. Through thoughtful aggregation of information and highlighting of important post-handoff events, preliminary results indicate that the system effectively reduces the barriers for providers to “close the loop” on prior clinical decisions. Our initial qualitative results are positive from residents, attending physicians and patient safety leaders. Through ongoing interviews with stakeholders, we are continuing to refine and expand the system.

This project is currently limited to a single academic institution however, we believe that our institution’s model with frequent handoffs of care is not unusual. Our system also currently only identifies patient handoffs for services which designate individual names on the “treatment team” in the EHR.

Future directions include: (1) further integration of the system into our EHR to enable more interactivity with the Learning List (e.g., dynamic addition/removal of patients to list, etc.); (2) quantifying the impact of the system on follow-up of patient outcomes; (3) enabling functionality to allow providers to flag cases for further learning (e.g., cases appropriate for teaching or morbidity & mortality conferences, cases which warrant further patient safety review, etc.); and (4) expansion of the system to support medical students and additional services.

Acknowledgement

This project is funded through a grant from the Gordon and Betty Moore Foundation.

References


1333
EHR Sequencing: A Novel Approach for Constructing Predictive and Interpretable Data Representations from EHR Data

Hossein Estiri, PhD\textsuperscript{1,2,3}, Thomas H. McCoy, MD\textsuperscript{1,2}, Shawn N. Murphy, MD, PhD\textsuperscript{1,2,3}  
\textsuperscript{1}Harvard Medical School; \textsuperscript{2}Mass. General Hospital; \textsuperscript{3}Partners Healthcare, Boston, MA

Introduction

Despite the natural excitement emerging from the large amount of information presented by electronic health records (EHRs), daunting methodological challenges remain for effective secondary use of EHR data. EHRs contain important temporal information about disease progression and patients that can be harnessed to discover important medical knowledge\textsuperscript{1,2}. However, the temporal nature of observational data stored in EHRs has not been fully exploited by current methods\textsuperscript{3}. The time-stamped EHR observations are often acquired asynchronously and exhibit issues of sparsity and heterogeneity, thus provide fundamental methodological challenges for applying common temporal analysis methods\textsuperscript{4,5}. Advanced learning algorithms such as Recurrent Neural Networks and its derivatives have been utilized to synthesize temporality of EHR data. These algorithms often result in highly predictive models, but they are hard to understand, limiting their application in healthcare settings. While interest in approaches to process temporal data in EHRs is growing, progress is inhibited by complexities of adequately representing time in EHR observations. Only recently have investigators begun developing and applying methods to extract temporal knowledge from EHR data. Clinical observations in EHRs are often stored at multiple temporal granularities, requiring transformation into a more uniform dimension for data mining. Properly incorporating temporality of EHR observations can present new signals and yield to interpretable findings from large scale clinical databases. Temporal abstraction methods use domain knowledge to transform multivariate temporal data into a set of symbolic interval-based presentation of the clinical observations captured in EHR data\textsuperscript{6}. Considering electronic health records as ‘indirect’ reflections of a patient’s true health state, we apply temporal abstraction by sequencing EHR observations to construct new EHR data representations that are both interpretable and highly predictive.

Methods

We propose a high-throughput procedure for constructing the new sequenced data representations and test their prediction and classification performance (Figure 1), using medications and diagnoses observations from 207 chart reviewed patients with gold standard labels for congestive heart failure (CHF). We first utilized the i2b2 hierarchical ontologies implemented in the Research Patient Data Registry (RPDR) from Partners HealthCare to construct a set of raw EHR features at different levels of the hierarchy. Second, we performed temporal abstraction by sequencing observations. We only utilized the first occurrence of an observation at a given time and constructed 2-deep sequences (time \(n\) and time \(n+1\)) for each patient. Third, we performed dimensionality reduction through correlation-based feature screening and embedded feature selection to identify a set of highly informative features. In the correlation-based analysis, we excluded sequences that did not have a statistically significant correlation with the CHF labels. The embedded feature selection involved using Random Forest classifiers to compute Mean Gini Decrease (MDG), as a metric for variable importance, and rank the top features based on their importance.

In the fourth phase, we fed the selected features to \(L_1\) logistic regression models in order to perform classification and prediction – for prediction, we removed observations from the date of the first record for CHF in the medical records. The experimental design involved 100 iterations of the classifiers. We compared the area under the receiver operating characteristic curve (AUC ROC) obtained from both raw and sequenced data representations at different ontological classes.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{The 4-step study design.}
\end{figure}
Results

Utilizing the full data for classification, we mined 1,013,052 unique 2-deep sequences (or sequenced data representations) from the 20,001 raw EHR observations, extracted from the 4D4M ontology class. The correlation-based screening reduced the dimensionality to 1,141 sequenced data representations (at $p<0.01$). Overall, we found that sequenced data representations significantly outperformed raw EHR features in classifying congestive heart failure, by more than 10 percent, depending on the ontology levels. Ranking the classifiers by ontology class, we found that using the sequenced data representations from the 4th level of ontology for both diagnoses and medications resulted in the best classifier performance (mean AUC ROC of 0.926). Moreover, we also found that the sequenced data representations presented novel insight that are difficult to discern from raw EHRs observations. Based on the classification results, we used the 4th level of ontology hierarchy for prediction, which revealed more information about the predictive power of the sequenced data representations. In comparing the AUC ROCs in predictive classifiers, we also examined different number of features (Figure 2).

![Figure 2. The sequenced data representations collectively outperform raw EHR data.](image)

Discussion

We proposed a new approach to constructing predictive and interpretable data representations from EHR data through temporal sequencing. The new sequenced data representations capture the temporality nature of EHR observations and is capable to encode multiple parameters into a unique data representation. For example, in the simplest form of 2-deep sequencing, parameters A and B (two unique observations), and the information that B happened after A (a temporal dimension) are encoded into a single data representation, $A \rightarrow B$. In addition to the dimensionality reduction this approach offers, our results have demonstrated that harnessing knowledge of disease progression through temporal sequencing improve computational disease classification (phenotyping) and prediction using EHR observations. We found that sequenced data representations carry more information about existence of the disease and are also useful for predicting the disease onset before an observation of the disease is recorded in the EHRs. The sequenced data representations present information beyond single clinical observations that can potentially shed light on some dimensions of the healthcare processes. Given the rapidly increasing prevalence of electronic health record systems in today’s practice, exploiting the temporal information in EHRs can advance medical knowledge discovery and meaningfully change clinical care by identifying and validating novel disease markers.

References

Growing Disparities in Use of Policy-Supported eHealth Tools

Jordan Everson, PhD, MPP\textsuperscript{1}, Evan Butler, BA\textsuperscript{1}, Nicole Senft, PhD\textsuperscript{1}

\textsuperscript{1}Vanderbilt University School of Medicine, Nashville, Tennessee

INTRODUCTION

Online access to health information and health services (eHealth) has increased along with internet use (1, 2). Public policies have sought to accelerate the spread of eHealth tools related to communication between patients and providers but have not targeted other eHealth tools, such as those facilitating access to online health information that is independent of healthcare providers. Since 2000, observers of the diffusion of eHealth tools across segments of the U.S. population have noted a “digital divide,” with members of traditionally underserved groups (lower SES, older, racial/ethnic minorities) being less likely to engage in many eHealth activities.(3-5) However, it is not clear whether the digital divide in the use of eHealth tools targeted by public policy has lessened or grown in the years since enactment of supportive policy compared to eHealth tools that have not been directly targeted by policymakers.

To address this gap, we describe use of one policy-supported technology (communicating with providers via messaging) and one technology outside the scope of public policy (looking for health information online) over time and across socioeconomic strata and racial groups.

METHODS

We used data from eight iterations of the National Cancer Institute’s Health Information National Trends Survey (HINTS). The full sample included 37,300 individual-year responses from 2003-2018. We measured use of provider messaging using an item asking participants whether, in the past 12 months, they had, “used e-mail or the Internet to communicate with a doctor or a doctor’s office.” Looking for health information online was assessed by an item asking whether, in the past 12 months, they had used the internet to “look for health or medical information for yourself.”

We defined socioeconomic strata by level of education, categorized as four levels in all years: less than high school, high school graduate or GED, some college or technical school, college graduate or greater. To measure race/ethnicity, we defined six different racial groups: Hispanic, White, Black, Asian, American Indian/Native Alaskan/Native Hawaiian/Pacific Islander, multiracial. We also defined several related demographic and health-related variables for inclusion in multivariate models.

We compared trends in looking for health information online and provider messaging by education and race/ethnicity. In multivariate analysis, we compared how education and race/ethnicity related to provider messaging and looking for health information online in three periods: before major public investment (2005-2007), the first years of public support (2011-2013) and recent years (2017-2018). Analyses adjusted for household income, sex, age, marital status, insurance coverage, and general health.

RESULTS

Overall Trends

Growth in provider messaging was relatively slow during the first years of the study period, increasing by 9.0 percentage points (from 4.4% to 13%) between 2003 and 2011 and was more rapid in later years, increasing by 22 percentage points between 2011 and 2018 (from 13% to 36%). In contrast to provider messaging, growth in looking for information online was rapid during the first years of the study period and slowed in later years: from 2003-2011 use grew by 31 percentage points (from 32% to 63%), then increased by only 7.2 percentage points between 2011 and 2018 (from 63% to 70%).

Trends in Disparities

Education

Among individuals who did not complete high school, reported rates of provider messaging increased by 13 percentage points between 2003 and 2018, the smallest increase of any education group. Meanwhile, among individuals who completed college, the rate of provider messaging increased by 46 percentage points over the same period, the largest increase of any education group. Compared to provider messaging, the rate of increase in looking for health information online was more similar across education levels: reported rates of looking for information increased by 36 percentage points for individuals without a high school degree and 33 percent among college graduates.
**Race/Ethnicity**

Among Hispanics, reported rates of provider messaging increased by 22 percentage points between 2003 and 2018, the lowest rate of any group. Meanwhile, among non-Hispanic Whites, the reported rates of provider messaging increased by 35 percentage points. In contrast to changes in provider messaging, reported rates of looking for health information online increased most quickly among traditionally underserved racial groups. Among Hispanics, reported rates of looking for health information online increased by 49 percentage points between 2003 and 2018, the highest rate of any group. Meanwhile among non-Hispanic Whites, the reported rates of looking for health information online increased by 38 percentage points.

**Adjusted Predictors of eHealth Use**

In multivariate models, the magnitude of the association between education and provider messaging increased over time but were smaller than unadjusted differences. In comparison, the adjusted association between education and looking for health information online grew less strong over time. As with education, we observed diverging trends between race/ethnicity and rates of provider messaging and looking for health information online. In adjusted models in 2003-2005, non-Hispanic Whites were only slightly more likely to use provider messaging than Hispanics, but this difference grew by 2017-2018. In contrast, the adjusted association between race/ethnicity and looking up health information online lessened over time.

**DISCUSSION**

In data from 2003 to 2018, the “digital divide” in use of public policy supported, provider-focused eHealth (provider messaging) has widened, while the divide in independent eHealth (looking up health information online) has stayed the same. While use of both these activities has increased markedly, rates of provider messaging grew more rapidly in the years following public policy intervention, while rates of looking up information online began to stagnate during the same period. These findings indicate that public policy may have impacted growth in the technologies it targeted, but may have disproportionately accelerated growth among White and well-educated individuals relative to socioeconomically disadvantaged groups. Future policy should seek to decrease this gap, perhaps by creating outreach efforts to encourage disadvantaged patients to use eHealth to communicate with providers.

**REFERENCES**

Tuning and Evaluating a State-of-the-Art Machine Comprehension Model for Clinical Why-Question Answering

Jungwei Fan, PhD\textsuperscript{1,2}
\textsuperscript{1}Division of Digital Health Sciences, \textsuperscript{2}Kern Center for the Science of Health Care Delivery
Mayo Clinic, Rochester, Minnesota

Introduction

The reasoning and decision-making in clinical practice can be naturally framed into a series of questions and answers. Making computers perform question-answering (QA) is deemed an accomplishment in artificial intelligence (AI) and vitally researched in the medical domain. Among the diverse information needs, why-QA is a distinct category that deals with cause, motivation, circumstance, and justification. In terms of prevalence, 20\% of the top ten question types asked by family doctors\textsuperscript{1} can actually be paraphrased into a why-question. Clinical why-QA is of great significance because: 1) toward a deeper solution the task resembles expert-level explanatory synthesis of knowledge and evidence, 2) toward a shallower solution it would practically identify decision reasons documented in clinical narratives.

The current study focuses on the second type of solution above, a modest yet very useful task of reason identification. It represents a “patient-specific” QA scenario in which a system tries to identify answers documented in the clinical notes. For example, why was his dobutamine stress test rescheduled? This is different from the other more commonly researched scenario of consulting knowledge sources to answer a non-patient-specific question. In non-medical domains, such patient-specific QA has its counterpart known as reading comprehension QA, with competitive open challenges and richly-annotated corpora. SQuAD 2.0\textsuperscript{2} is an iconic corpus and challenge, which requires QA systems to refrain from answering when no suitable answer is present in the reference document. A modeling approach that caught wide attention recently was Google’s BERT,\textsuperscript{3} for it achieved impressive performance not only in SQuAD 2.0 but multiple AI challenges. BERT offers two pre-trained language models for task-specific fine-tuning: BERT\textsubscript{base} (12 layers 110M parameters) and BERT\textsubscript{large} (24 layers 340M parameters). Another notable resource is emrQA,\textsuperscript{4} a large clinical QA corpus generated from the i2b2 NLP challenge datasets and with about 7.5\% of why-QA annotations.

As an initial step toward developing a clinical reason identification system, this study aims to adapt the promising BERT models for why-QA based on the patient’s notes. The experiments use the why-subset of emrQA for training/testing and require the system to execute judicious refraining as in the SQuAD 2.0 challenge. Quantitative and qualitative evaluations are performed to assess model capability and inform future improvement.

Methods

A subset of 27,517 QA pairs were extracted from emrQA by requiring that the question starts with “why”. Also, to train a why-QA model that can refrain from giving an answer when none exists in the text, an additional 2,839 negative why-QA pairs were automatically generated where neither the question anchor nor the answer presented in the clinical note, i.e., an unanswerable instance. Together the positive and negative instances formed the core study set, emrQA\textsubscript{why}. The emrQA\textsubscript{why} was then split into train/dev/test subsets of 13,083/4,203/13,070 instances respectively. The dev subset mainly served to tune the optimal cutoff for making the do-not-answer decision. Four models were trained and evaluated for comparison: 1) BERT\textsubscript{base} fine-tuned by emrQA\textsubscript{why}, 2) BERT\textsubscript{large} fined-tuned by emrQA\textsubscript{why}, 3) BERT\textsubscript{base} fine-tuned first by 1,833 SQuAD 2.0 why-QAs and then by emrQA\textsubscript{why}, and 4) BERT\textsubscript{base} first pre-trained by 1,474 independent notes (106,952 sentences) from the i2b2 NLP corpora and then fine-tuned by emrQA\textsubscript{why}. The fine-tunings of BERT\textsubscript{base} used 4 epochs with –max_seq_length=128 and –batch_train_size=32, while in fine-tuning BERT\textsubscript{large} –batch_train_size=6 was used to avoid out-of-memory. The jobs were run on a Tesla K80 with compute capability 3.7 and 11GB of memory. The evaluation followed standard SQuAD metrics, comparing system answers to the gold by exact and partial match (average F1 between the predicted and gold bags of tokens).

Results and discussion

(please refer to Table 1, Table 2, and Figure 1)

The total time used in training the four models in sequential order as shown in the tables were: 1h 49m, 5h 43m, 1h 56m, and 48h 27m respectively. As recommended by the official manual, –batch_train_size=6 was used in fine-tuning BERT\textsubscript{large} to avoid memory issue. However, it resulted in suboptimal accuracy due to the inadequate batch size, as warned in the manual. The injection of SQuAD\textsubscript{why} (in the third model) was to assess the smoothing effect from non-medical training data. Interestingly, the out-of-domain why-instances caused worse accuracy on the dev set but better accuracy on the larger test set, suggesting mild supplemental benefit beyond merely BERT\textsubscript{base} + emrQA\textsubscript{why}. Extensive re-training of BERT\textsubscript{base} with additional i2b2 notes was time-consuming (almost 2 days) but most helpful (~3\% increase in accuracy on the larger test set).
of accuracy). This indicates the importance of domain-customizing the fundamental language model. Although the system errs toward the conservative side (heavy refraining and recall capped at ~0.6 as shown in Figure 1), the predicted answers above the cutoff tend to be precise around 0.8, which is desirable for the target application. Model inspection and error analysis are ongoing with consulting domain experts. The preliminary inspection suggests that some predicted answers may be semantically acceptable, even not literally matching the gold annotation.

Table 1. Accuracy of differently tuned models on the dev (development) set of 4,203 QA pairs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Full dev set: 4,203 QAs</th>
<th>Dev HasAns: 3,764 QAs</th>
<th>Dev NoAns: 439 QAs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>F1</td>
<td>Exact</td>
</tr>
<tr>
<td>BERT\textsubscript{base} + emrQA\textsubscript{why}</td>
<td>0.673</td>
<td>0.724</td>
<td>0.636</td>
</tr>
<tr>
<td>BERT\textsubscript{large} + emrQA\textsubscript{why}</td>
<td>0.650</td>
<td>0.691</td>
<td>0.610</td>
</tr>
<tr>
<td>BERT\textsubscript{base} + SQuAD\textsubscript{why} + emrQA\textsubscript{why}</td>
<td>0.662</td>
<td>0.716</td>
<td>0.624</td>
</tr>
<tr>
<td>BERT\textsubscript{base} + Pretraining\textsubscript{i2b2} + emrQA\textsubscript{why}</td>
<td>0.700</td>
<td>0.752</td>
<td>0.667</td>
</tr>
</tbody>
</table>

Table 2. Accuracy of differently tuned models on the test set of 13,070 QA pairs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Full test set: 13,070 QAs</th>
<th>Test HasAns: 11,944 QAs</th>
<th>Test NoAns: 1,126 QAs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>F1</td>
<td>Exact</td>
</tr>
<tr>
<td>BERT\textsubscript{base} + emrQA\textsubscript{why}</td>
<td>0.624</td>
<td>0.686</td>
<td>0.589</td>
</tr>
<tr>
<td>BERT\textsubscript{large} + emrQA\textsubscript{why}</td>
<td>0.605</td>
<td>0.661</td>
<td>0.568</td>
</tr>
<tr>
<td>BERT\textsubscript{base} + SQuAD\textsubscript{why} + emrQA\textsubscript{why}</td>
<td>0.632</td>
<td>0.693</td>
<td>0.597</td>
</tr>
<tr>
<td>BERT\textsubscript{base} + Pretraining\textsubscript{i2b2} + emrQA\textsubscript{why}</td>
<td>0.656</td>
<td>0.710</td>
<td>0.625</td>
</tr>
</tbody>
</table>

Figure 1. Precision-recall curves of the BERT\textsubscript{base} + Pretraining\textsubscript{i2b2} + emrQA\textsubscript{why} model evaluated on the test set

Conclusion

The state-of-the-art BERT language model is tuned and evaluated for the task of clinical why-QA. The results show promising precision and indicate the importance of fundamental domain customization. The qualitative analysis should shed light on model behavior and areas to improve. The main deliverable will be a reason identification system. The study did not compare with other baseline models, which is a limitation to be addressed in future work.

References

Tracking AMIA Health Informatics Educational Domains:
Lessons Learned from a Pilot Test

Sue S. Feldman, RN, MEd, PhD1, Sue Boren, PhD2, David E. Moxley, MLS2
Courtney B. Garza, MS1, Eta S. Berner, EdD1
1University of Alabama at Birmingham, Birmingham, AL, USA,2University of Missouri, Columbia, MO, USA

Abstract

This presentation will discuss different approaches for using Canvas, a learning management system, to track competencies mapped to AMIA Health Informatics foundational domains. This presentation extends the work of one organization (UAB), presented at the AMIA Academic Forum, to include the University of Missouri. Each program will discuss how they went about setting up Canvas to be used in a novel way. The presentation will include a description of data captured, reports produced, and how the data and reports are used to inform course and program improvement and accreditation activities. Lessons learned from a pilot test will also be presented.

Introduction

AMIA has published foundational domains to support graduate Health Informatics education at the Masters level. These foundational domains provide a framework for understanding the degree to which individual Health Informatics programs are addressing each foundational domain and at what level of competence. Competencies, as defined in an earlier AMIA document, are the integration of knowledge, skills, and attitudes in F4-F10 in Error! Reference source not found.; F1-F3 are working knowledge only without skills or attitudes attached (1).

Health Informatics programs will use these foundational domains to support competencies such that they can be measured and assessed. In understanding the relationship between foundational domains and program specific competencies, it is important to understand that all 10 foundational domains must have a degree of integration dependent on the focus of the individual program.

Methods

Repurposing existing Canvas fields, we were able to collect data and produce reports to inform course and program
improvement, AMIA foundational domain areas, and competencies. We then conducted a pilot test with two courses. The process started with the program competencies. Next, each course objective within each course used in the pilot was assessed to ensure alignment of the program competency. In other words, understanding if the course objectives align to program competencies align to program competencies. Then each course objective was aligned with the assignments and assessments used to measure that objective. The assignments and assessments were also aligned to the AMIA foundational domains. For each objective, the collective of the assignment, the assessment, and the foundational domain were then aligned to the appropriate level on Miller’s Pyramid. Figure 3 illustrates this process.

After this process was completed, we utilized the Outcomes feature in Canvas. The “outcome” in Canvas indicates the domain, the competency, and the level of Miller’s pyramid for each aligned assessment. Each domain/competency has a rubric with the levels of Miller’s pyramid indicated. These rubrics were added to the aligned assignment/assessment so that faculty could indicate, under the appropriate pyramid level, whether the outcome (domain/competency) was achieved in that assignment. When the assessment was accomplished at 80% or better, the competency was said to be attained and faculty scored that outcome as a 1 under the appropriate Miller’s level. If a student achieved less than 80% they were given a 0 on the outcome score. From these data, individual student and aggregate reports are able to be produced and analyzed.

Results

The results across the two university Health Informatics programs were two fold: 1) at a system level, using an existing LMS in a novel manner and 2) at a functional level, data reporting. Most LMS’s were intended more as a vehicle to store and disseminate course information. At a system level, we found that the ability to use the LMS to track and assess the AMIA foundational domains and student competencies meant that we did not need to develop another tracking program. Secondly, the outcome results report, which includes data for both courses, displays which students achieved which competency and at which Miller’s Level at the individual course level and across the program. These data are then aggregated and used to inform course and program improvement as well as accreditation activities.

Discussion of Results

Most LMS’s contain analytics centered around usage and grading. By repurposing some of the back-end administrative fields in the LMS, we were able to pull data out of the LMS that is better targeted at tracking and assessing the AMIA foundational domains and student competencies. The process, at this point, is a step in the right direction; however some challenges remain. Rather than have Canvas automatically apply a “score” based on a grade of 80%, the instructor needs to use the competency rubric at the point of grading to indicate attainment.

Conclusion

Integrating curriculum maps into Canvas facilitates monitoring achievement and measuring outcomes relative to program competencies. Doing so allows for a continuous process of program self-assessment. From this small pilot study, we conclude that learning management systems can be used for program monitoring. However, this small pilot illuminates an opportunity to work with LMS developers so that they better understand program needs. Doing so could create a “competency module” for use across multiple programs and broaden the use of a LMS in actionable data collection. Absent such a module, programs should examine their learning management system’s capabilities as to its ability to facilitate data collection that can be used to document outcomes for accreditation purposes. More exploration of this process is needed and additional reporting of future attempts at this or a similar project would inform the use of an LMS for competency capture and reporting.

References

Towards machine-learning informed early detection of dementia in UK primary care

Elizabeth Ford, MA DPhil¹, Johannes Starlinger, MD PhD²
¹Brighton and Sussex Medical School, Brighton, UK; ²Charité – Universitätsmedizin Berlin, Berlin, Germany

Abstract
We used UK primary care EHR data to discover a minimum set of best features which could identify the onset of dementia. After specifying 88 possible predictive features, we used random forest classifiers to identify the top features in the 5 years prior to diagnosis, in a dataset of >90,000 dementia cases and matched controls. Classification performance improved significantly within the last year before diagnosis, when memory loss was evidently recognized by clinicians.

Introduction
Dementia is a global public health challenge and is one of the most common and serious disorders in the elderly population. Dementia is characterized by progressive decline in memory and cognition; symptoms can be caused by a number of illnesses which lead to structural and chemical changes in the brain, and death of brain tissue¹. Currently there are no treatments which are curative for dementia, however, a timely diagnosis of dementia can still be valuable for patients, as they can be offered a range of support interventions, inform themselves about the condition, and plan for the future, which may help them maximize their quality of life². Yet, at the current time, only two-thirds of patients with dementia receive a diagnosis for the condition in the United Kingdom National Health Service³.

Primary care physicians, known in the UK as general practitioners (GPs), record information about all interactions with their patients in comprehensive electronic health records (EHRs). These records provide a valuable resource for understanding the onset and early presentation of dementia as recorded in general practice, and may provide a rich data source for creating diagnostic support systems for GPs to recognize dementia in a more timely way.

Many previous studies have modelled risk for dementia at a future time, but no models exist to detect patients who are in the early stages of dementia but as yet undiagnosed, using only clinical primary care data. The objective of this study was to use machine-learning methods to inform a minimum set of best features for dementia detection within the primary care EHR, and to investigate how early the onset of dementia could be detected in this dataset.

Methods
This study used data from the UK Clinical Practice Research Datalink (CPRD), which curates anonymized EHRs from UK primary care, from 13 million patients⁴. Data are captured using a structured hierarchical vocabulary called Read Codes; each of the 200,000 7-byte alphanumeric codes maps onto a term or short phrase describing a clinical concept. For this project we extracted the entire EHR of all patients who received a diagnosis code for dementia between 2000 and 2012 (N=47,653) and the entire EHR for a similar number of controls, matched on age, sex and GP practice and with no dementia code anywhere in their record (N=47,127).

Our feature selection was informed by a meta-analysis⁵ and a survey of GPs. This generated 88 separate clinical concepts, including conditions, symptoms, medications and other factors, which were thought to precede dementia diagnosis in the EHR; these were operationalized into Read Code lists.

We partitioned the original CPRD dataset to represent the course of care over time: the first partition (-5 years) holds all data within the fifth year prior the date of a patient’s initial diagnosis of dementia; the second partition (-4 years) includes the fifth and fourth year prior to diagnosis; and so forth. The last partition (-1 years) contains all data within the whole 5-year window up to 1 day before the data of diagnosis (excluding the first appearance of an actual dementia code). Matched controls were analogously included in each partition. For these, the index date is determined by the matching process. Within each of these partitions, we vectorized patients and controls by mapping the Read Codes recorded in the data to the feature space of the 88 code lists of relevant clinical concepts. A binary feature representation was used to model a simple, checklist style symptom assessment in practice.

We trained Random Forest classifiers for each partition with holdout test sets of 33%. The random forest algorithm was chosen because it intrinsically includes feature weighting and selection, which we exploited to identify the most
important features indicative of an upcoming diagnosis.

Results

After data cleaning, 93120 patients were included in the final analysis, of whom 35% were men and 65% were women. The median age at diagnosis date (or matched index date) was 82.6 years (range 64.5-109.9 years). All patients had at least three years’ worth of data, 90,351 patients (97.0%) had ≥ four years and 87,876 patients (94.4%) ≥ five years.

Figure 1 shows the ROC curves of the resulting models for each partition. In investigating how early detection could be achieved, we noted that classification performance increased significantly within the last year before diagnosis, preceded by only slight but continuous increases from years -5 to -2. Inspecting the five most relevant features within each partition revealed that this increase in model performance correlates to a gradual shift in the types of codes recorded for patients over time (see Table 1): while 5 years prior to diagnosis long term risk factors are predominant, the importance of codes related to health events and to symptoms of memory loss continuously increases towards the point of diagnosis. In the last year preceding diagnosis, action towards diagnostic underpinning becomes apparent.

Discussion

The observed change of features which contribute most strongly to the predictive classifiers, together with the continuous increase in classification performance, indicates that identifying onset of the dementia using decision support, much before GPs pick up symptoms, may be problematic. Future research on early detection of dementia could focus on linking multiple data sources, such as patient wearables, in order to gain insights on disease status which clinicians do not have access to. However, our list of best features gives clear indications of patients most at risk, and includes dementia risk factors, such as smoking and diabetes which are well-established in the literature. It is notable that memory loss symptoms are recorded but action towards a formal dementia diagnosis may not be taken for a number of years, this may be due to stigma or patient unwillingness. Our results suggest that delay in diagnosis may be related to patient or system factors rather than symptoms going undetected by physicians. We plan to work towards a feature set which can be operationalized into a computerized risk communication and discussion tool for GPs to use with their high risk patients.

References

Introduction

Understanding how effectively healthcare providers use Electronic Health Records (EHRs) to find relevant patient information is an active area of clinical informatics research. In the primary care setting, prior work has used cognitive task analysis to understand challenges in the efficiency and comprehensiveness of chart review. However, it is unclear, whether specific categories of search patterns exist, and, if so, whether these patterns are typical to a provider’s use of the EHR or are related to a focus of disease management. In our work, we used grounded theory to identify distinct categories of search used by primary care physicians (PCPs) reviewing charts of patients with diabetes and/or hypertension. We present a cognitive workflow model that can visually illustrate PCP information seeking patterns. We discuss the implications of our study results on EHR interface design and future work.

Method

Eight PCPs recruited from Atrius Health (a medical practice group in Massachusetts) consented to participate in the IRB-approved study. Ten patients, who were not under the care of the PCPs, consented to have their EHR records used for this study. The patients were over 18, had been seen at the medical practice for over 10 years, and had at least one active condition of hypertension and/or diabetes. Each PCP reviewed two to four records in a quasi-random presentation order. For each record, the physician was asked to see the patient for the first time for a chronic disease management follow-up. The session was audio-recorded.

Twenty-seven audio-recordings were collected and manually transcribed. Themes were extracted and encoded by the authors MF, YP, JS and SM using grounded theory. To understand what information is required and why that information is needed, we devised a coding system consisting of Information Seeking Categories (ISCs) and Intents. During the coding process, discrete segments of each transcript were marked as a unit of meaning that included both an ISC and a corresponding intent. During the initial pass, two authors independently reviewed and coded each transcript. Any disagreements were reviewed by the larger group, so that full intercoder agreement was achieved.

We used a network representation (Figure 2) to visualize the PCP’s cognitive workflow between discrete ISC segments during the chart review. In the network model, the nodes (or vertices) represent the various ISCs and a directed link (or edge) between nodes shows how the physician moved from one ISC to another ISC (or the same ISC) in the following segments found in the transcript. The size of a node represents the number of times that a single ISC was coded in the record review, and the width of the link indicates the frequency of times the PCP’s review transitioned from one ISC node to another ISC node (or the same node).

Results

Our coding process discovered nine types of ISCs. The ISCs of “Problem List”, “Medications”, and “Labs and Vitals” relate to the central elements of an EHR record. “Person’s Story” involves looking for social determinants of health. The ISCs of “Tests and Imaging” and “Surgeries” correspond to search related to these procedures. “Health Maintenance” includes search for routine procedures and tests. “Specialist Notes” and “Care Team” capture searches for information about who else treated the patient.

The average number of ISCs generated per physician per chart review was 51.3 (standard deviation 39.6). Figure 1 shows the average frequency of ISCs across all chart reviews. The top four ISCs in chart review, “Lab and Vitals”, “Problem List”, “Person’s Story” and “Medications”, accounted for over 80% of the information search. We measured the Markov probability of transition from one ISC to another ISC across all chart reviews, which we present as a transition matrix (Table). As shown visually by the color intensity, search of most ISCs was followed by further search of the same ISC. Otherwise, many types of searches were frequently followed by searches of the four most common ISCs. In inspecting the network representations of the cognitive workflows in the 27 chart reviews, we found a wide variety of search patterns. Figure 2 shows the different types of search patterns in the cognitive workflow of two PCPs reviewing the same two patient charts and illustrates that the pattern is neither specific to the PCP nor to the patient.
Table. Probability of Transition from one Information Seeking Category (ISC) to another or same ISC

<table>
<thead>
<tr>
<th>From ISC</th>
<th>Labs &amp; Vitals</th>
<th>Problem List</th>
<th>Person’s Story</th>
<th>Medications</th>
<th>Health Maint.</th>
<th>Care Team</th>
<th>Tests &amp; Imaging</th>
<th>Specialist</th>
<th>Surgeries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labs &amp; Vitals</td>
<td>0.51</td>
<td>0.12</td>
<td>0.10</td>
<td>0.15</td>
<td>0.05</td>
<td>0.04</td>
<td>0.02</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Problem List</td>
<td>0.18</td>
<td>0.38</td>
<td>0.12</td>
<td>0.10</td>
<td>0.07</td>
<td>0.07</td>
<td>0.04</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>Person’s Story</td>
<td>0.14</td>
<td>0.22</td>
<td>0.36</td>
<td>0.14</td>
<td>0.06</td>
<td>0.04</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Medications</td>
<td>0.23</td>
<td>0.18</td>
<td>0.09</td>
<td>0.39</td>
<td>0.07</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Health Maint.</td>
<td>0.14</td>
<td>0.09</td>
<td>0.15</td>
<td>0.14</td>
<td>0.36</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Care Team</td>
<td>0.16</td>
<td>0.21</td>
<td>0.18</td>
<td>0.00</td>
<td>0.09</td>
<td>0.28</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Tests &amp; Imaging</td>
<td>0.02</td>
<td>0.24</td>
<td>0.12</td>
<td>0.05</td>
<td>0.02</td>
<td>0.07</td>
<td>0.39</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>Specialist</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.17</td>
<td>0.25</td>
<td>0.00</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>Surgeries</td>
<td>0.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Figure 2. Cognitive workflows used by two primary care physicians (PCP1, PCP2) reviewing two patients (PT1, PT2). Abbreviations used: HM (Health Maintenance), LV (Labs and Vitals), Meds (Medications) and PL (Problem List).

Discussion

Our macro-level analyses indicate that, as expected, PCPs search for common ISCs relevant to chronic disease management. However, we also found significant variation in PCP foraging patterns at the micro level for each patient, as illustrated in the cognitive workflow visualizations. These four examples, plus the other twenty-three workflow diagrams, do not reveal prototypical search patterns indicative of a particular PCP’s style of chart review or of a focused disease management activity (such as achieving a treatment goal). These results suggest that efforts to redesign EHR interfaces to support linear workflows may inhibit the exploratory nature of PCP chart review. To reduce information foraging, our results suggest that frequently searched ISCs should be kept close together whereas less frequently needed ISCs can be further apart. In future work, we plan to analyze categories of intents for ISCs and examine the hierarchical nature of search behaviors. We will also assess whether our network-based representation for modeling cognitive workflow is broadly informative for understanding mental models in team-based care.

References

Experiences from the National Demonstrator Study within the German Medical Informatics Initiative

Thomas Ganslandt, MD1, Jannik Schaaf, MSc2, Josef Schepers3, Holger Storf2, Felix Balzer, MD3, Silke Haferkamp, MD4, Robert Lodahl, MSc5, Fabian Prasser, PhD6, Ulrich Sax, PhD5, Holger Stenzhorn7, Hans-Ulrich Prokosch, PhD8, Martin Boeker, MD9

1 Department of Biomedical Informatics of the Heinrich-Lanz-Center, Mannheim University Medicine, Ruprecht-Karls-University Heidelberg, Mannheim, Germany; 2 Medical Informatics Group, University Hospital Frankfurt, Frankfurt am Main, Germany; 3 Department of Anaesthesiology and Intensive Care Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany; 4 Department of Information Technology, University Hospital Aachen, Germany; 5 Department of Medical Informatics, University Medical Center Göttingen. Göttingen, Germany; 6 Institute of Medical Informatics, Statistics and Epidemiology University Hospital rechts der Isar, Technical University of Munich, Munich, Germany; 7 Department of Information Technology and Applied Medical Informatics, University Hospital Tübingen, Germany; 8 Chair of Medical Informatics, Friedrich-Alexander-University Erlangen-Nuremberg, Germany; 9 Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center - University of Freiburg, Freiburg, Germany

Introduction. The German Medical Informatics Initiative (MII) is a large-scale, long-term strategic funding program by the German Federal Ministry of Education and Research to establish a nationwide infrastructure for the secondary use of routine clinical data at academic hospitals (1). While the initiative envisions a multi-year timeline for the four funded consortia to install interoperable data structures, terminologies and interfaces as well as to coordinate legal and ethical issues, it is imperative to continuously demonstrate and communicate successful collaboration throughout the project duration. Harvesting “low-hanging fruit” early on in a “demonstrator study” requires a fast-track approach that skirts longer-term efforts required to integrate the individual platform concepts pursued by the consortia but at the same time proves successful procurement of local data sets and the required legal and ethical permissions to analyze them.

Methods. A modular core dataset was designed in the cross-consortia Interoperability Working Group, which supports a stepwise implementation starting with a mandatory basic set of billing data modules (demographics, inpatient cases, diagnoses, procedures) supplemented by laboratory findings and medication modules as well as use-case specific extension modules (2). Inpatient billing data can be provided by all German hospitals in a standardized “§21” format mandated by the German Diagnosis-Related-Group-based billing regulations. An Extract/Transform/Load (ETL) process to import §21 conformant datasets into the i2b2 (Informatics for Integrating Biology and the Bedside) biomedical query platform (3) was implemented previously in the Integrated Data Repository Toolkit (IDRT) project and readily available for use (4). Two research topics were selected, which can be studied with §21 conformant data by combining locally aggregated analysis results at each participating site to guarantee patient data confidentiality: Comorbidity scores were calculated from ICD10 diagnosis data and analyzed in relation to age, duration of stay, type of hospital discharge and other variables. Additionally, rare diseases were analyzed based on ICD10 diagnosis data to generate geo visualizations and calculate distances from patient zip areas to treating centers. Based on the MIRACUM data protection concept (5), distributed analyses with suppression of local result sets with less than 5 patients were used to prevent re-identification of individual patients. Additionally, rare disease diagnosis codes were collated into larger disease groups. All analyses were centrally designed as SQL queries against the i2b2 PostgreSQL database, distributed to participating sites via a collaboration portal and executed locally. Aggregated anonymous result datasets were returned to the study coordinators for further analysis and visualization by means of the R statistics package. For the rare disease dataset, additional preprocessing was carried out locally at the sites using Python and R scripts. A study protocol was designed and provided to all participating sites to initiate approval through their respective ethics boards, data protection officers and use-and-access committees (UACs).

Results. 20 of 24 funded MII university hospitals chose to participate in the demonstrator study. The demonstrator platform was deployed to the sites as a pre-configured Linux virtual machine (VM) containing a dockerized i2b2 installation, PostgreSQL database and the IDRT ETL process for §21 conformant source datasets (Fig. 1). As of submission of this abstract, all 20 sites have internally procured datasets, deployed the demonstrator VM and received UAC approval, 19 sites have received ethics board and data protection approval, and 17 sites have provided aggregated result datasets. Based on these sites, a total of 1.7 million patients with 3.1 million inpatient cases (2015-
Discussion. The study successfully demonstrated active collaboration across MII consortia as well as the ability to procure datasets and gain permission to use them in the scope of a shared analysis. Preliminary results are consistent with expectations e.g. regarding the correlation between comorbidity scores and hospital discharge type. The use of standard datasets (§21 conformant) and established open source tools (i2b2, IDRT) was essential to facilitate the rapid implementation and deployment of the demonstrator platform in a timeframe of about 5 months.

Limitations: Even though meta-analyses of locally aggregated results were chosen to ensure patient privacy, the duration to acquire ethics board and data protection approval varied substantially between sites (also due to the different privacy regulations at the German state level). In this context, the demonstrator study may act as a first step towards streamlining the approval process for future similar projects. While the use of standardized billing data served to facilitate a quick entry into collaborative analyses, quality issues as well as the narrow scope of billing data need to be taken into account. Also, coverage of rare diseases in ICD10 is severely limited. Potential future iterations of the MII demonstrator study need to include additional core dataset modules (e.g. laboratory findings), systematically analyze data quality issues and, in the end, make use of the full arsenal of cross-consortia semantic harmonization, standardized data structures and legal coordination goals, as they are implemented throughout the MII funding period.

Conclusions. Standardized datasets and open source platforms were successfully leveraged to rapidly implement and deploy a demonstrator study for the investigation of the prevalence of comorbidities and rare diseases with clinical routine data within the German Medical Informatics Initiative. The approach is not limited to Germany, but could be leveraged for pilot projects in any environment that provides a shared dataset format (e.g. billing data).

Acknowledgements. This work was funded in part by the German Federal Ministry of Education and Research within the Medical Informatics Initiative, Grant IDs 01ZZ1801E (MIRACUM), 01ZZ1804A (DIFUTURE), 01ZZ1802A (HiGHmed), 01ZZ1609A (SMITH).

References
An Effort to A.C.T. (Accuracy, Comprehensiveness, and Timeliness) on Opioid Prescription Data

Esteban Gershanik, MD, MPH, MMSc, Brigham and Women’s Hospital, Boston, MA 02115, USA

**Learning Objectives:**
1. Understand the current manner in which opioid prescription rates are reported by the CDC and potential limitations for your state
2. Learn the benefits of and how to use the Prescription Monitoring Program (PMP) for more accurate, comprehensive, and timely data for addressing the opioid epidemic in your state.

**Introduction**
The opioid epidemic is a public health emergency in the USA with opioid prescriptions a contributing factor.1,2 To combat it, better data is needed for targeted interventions.1,3,4 The objective of this study is to compare the traditional CDC reporting of opioid prescription rates to the newer Prescription Monitoring Program (PMP) in Louisiana (LA) in hopes of more accurate, comprehensive, and timely data for targeted solutions.

**Methods:**
This study compares the LA opioid prescription rate reporting from the CDC and the state’s PMP. CDC uses the Quintiles IMS Data Warehouse that captures 88-94% of retail prescriptions in the USA and does not account for all counties. The PMP captures all state parish opioid prescriptions. Population data was captured from the USA Census Bureau. Trends were evaluated between the CDC reported USA and LA opioid prescription rates to that of the PMP from 2013 to 2016. For 2016, opioid prescription data was also broken down by each of LA’s 64 parishes between the two databases and main outcome measurements evaluated the gross and percentage difference between the two data sources.

**Results**
This study compares the traditional CDC reporting of opioid prescription rates to the Prescription Monitoring Program in Louisiana. While CDC and PMP both show a downward trend in LA opioid prescription rates from 2013 to 2016, there is a difference of an average 15 prescriptions between the two databases during those years (Figure 1). In 2016, database comparison demonstrated a 16% difference in opioid prescription rates reporting for the entire state (Figure 1) and greater than 100% discrepancy in 11 of 64 (17%) of Louisiana’s counties (Table 1).

**Discussion**
National reports and agencies have preached that better data is necessary to combat the exponential growth of the opioid epidemic.1,3,4 PMPs have shown to provide more accurate, comprehensive, and timely data on opioid prescriptions compared to previously used datasets and need to be utilized and optimized to ensure appropriately data-driven, targeted interventions in curbing this epidemic.
Illustrations:
Figure 1: CDC vs. PMP opioid prescription rate per 100 resident, 2013-2016

Table 1: Top CDC vs. PMP Louisiana Parish Prescription Rate Discrepancies (per 100 citizens), 2016

<table>
<thead>
<tr>
<th>Parish</th>
<th>Population Estimate*</th>
<th>CDC Prescription Rate</th>
<th>PMP Prescription Rate</th>
<th>Gross Difference (PMP vs CDC)</th>
<th>Percent Difference (PMP vs CDC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plaquemines Parish</td>
<td>23,464</td>
<td>47.6</td>
<td>105.20</td>
<td>57.60</td>
<td>121.01</td>
</tr>
<tr>
<td>Bienville Parish</td>
<td>13,865</td>
<td>58.4</td>
<td>131.50</td>
<td>73.10</td>
<td>121.37</td>
</tr>
<tr>
<td>Tensas Parish</td>
<td>4,597</td>
<td>44.6</td>
<td>103.55</td>
<td>58.95</td>
<td>133.17</td>
</tr>
<tr>
<td>West Carroll Parish</td>
<td>11,114</td>
<td>50.7</td>
<td>129.84</td>
<td>79.14</td>
<td>156.03</td>
</tr>
<tr>
<td>St. Helena Parish</td>
<td>10,512</td>
<td>73.1</td>
<td>197.05</td>
<td>123.96</td>
<td>169.58</td>
</tr>
<tr>
<td>East Feliciana Parish</td>
<td>16,883</td>
<td>46.6</td>
<td>133.47</td>
<td>86.87</td>
<td>186.42</td>
</tr>
<tr>
<td>Caddo Parish</td>
<td>10,087</td>
<td>51.8</td>
<td>154.48</td>
<td>102.68</td>
<td>198.22</td>
</tr>
<tr>
<td>Assumption Parish</td>
<td>22,695</td>
<td>33.1</td>
<td>106.20</td>
<td>73.10</td>
<td>220.86</td>
</tr>
<tr>
<td>Grant Parish</td>
<td>22,365</td>
<td>43.2</td>
<td>141.79</td>
<td>98.59</td>
<td>228.22</td>
</tr>
<tr>
<td>Madison Parish</td>
<td>11,512</td>
<td>1.8</td>
<td>78.44</td>
<td>76.64</td>
<td>425.51</td>
</tr>
<tr>
<td>Cameron Parish</td>
<td>6,882</td>
<td>–</td>
<td>95.41</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Population Estimate from the US Census Bureau

References:

The Role of Informatics in Social Needs Care Integration: Selected Results from a National Academies of Sciences, Engineering, and Medicine Committee Report

Rachel Gold, PhD, MPH1,2, Amber Haley, MPH3, Lisa A. Cooper, MD, MPH4, Laura Gottlieb, MD, MPH5

1OCHIN, Inc., Portland, OR; 2Kaiser Permanente NW Center for Health Research, Portland, OR; 3University of North Carolina, Chapel Hill, NC; 4Johns Hopkins University, Baltimore, MD; 5University of California, San Francisco, CA

Introduction: We present the final report of a National Academies of Sciences, Engineering, and Medicine (NASEM) ad hoc committee that was convened in July 2018, and charged with examining the potential for integrating 'social needs care' – e.g., services addressing social needs / social determinants – into the delivery of health care, to improve health outcomes. The committee’s report discusses current and emerging approaches to social needs care integration in diverse healthcare providers / systems; roles of different disciplines, care organizations, and types of providers; and efforts to inform the delivery of effective, efficient care that addresses social needs, improves health, and reduces health inequities. The report includes recommendations on how to: expand social needs care services; coordinate social needs care providers within interprofessional care teams; optimize the effectiveness of social services to improve health; integrate traditionally social and medical services; train and retain the workforce needed to support social needs integration; finance such integration; and address barriers to implementing this integration, including barriers related to technological and informatics. This presentation will touch on all aspects of the report, with primary emphasis on informatics-related barriers and facilitators to social needs integration into medical care. Committee members will be recognized in this presentation; they include experts in health services research, health equity, social work, etc.

Methods: A broad literature search was conducted. This was not a systematic review as the committee recognized that much relevant information would come from sources other than the peer-reviewed literature – e.g., the ‘grey’ literature such as reports and case studies. The search identified several thousand potentially relevant references, the titles and abstracts of which were reviewed. Of these, several hundred full-text articles were selected and reviewed, and a subset of them provided the evidence base used to support the report’s conclusions and recommendations. Three public sessions were held to obtain further information and perspectives, and the committee met five times in closed session to deliberate on the evidence and to develop conclusions and recommendations. The final report underwent exhaustive peer review.

Results: The committee report highlights the health information and technology needs of social and medical care integration, among other topics. It includes an extensive review and evaluation of informatics-related barriers associated with the implementation of electronic social risk documentation tools, and suggests potential facilitators to those barriers. The report addresses issues such as usability of the associated informatics tools, and associated staff training and resource needs, interoperability concerns, and data safety and patient privacy concerns. Details from the report will be given at this presentation, but cannot be listed here as the report is not yet public.

Discussion: The NASEM committee report could influence policies for health information technology innovations related to medical and social care integration; if feasible, the presentation will discuss how the report has been used since its release (September 2019). Committee recommendations about how to design and integrate technologies that better support social and medical care integration are likely to influence adoption in new settings.

References

Interventions to Increase Patient Portal Use in Vulnerable Populations: A Systematic Review

Lisa V. Grossman¹, Ruth M Masterson Creber², Natalie C Benda², Drew Wright³, David K Vawdrey¹, Jessica S Ancker²

¹ Department of Biomedical Informatics, College of Physicians and Surgeons, Columbia University, New York, NY, USA; ² Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, NY, USA; ³ Samuel J Wood Library, Information Technologies and Services, Weill Cornell Medicine, New York, NY, USA

Introduction

Last year, millions of Americans accessed their own health records online, more than ever before.¹ Secure websites called patient portals offer convenient, 24-hour access to records, as well as other health management features. Portals provide patients with unprecedented transparency into health information, which evidence suggests may prevent medical errors, increase shared decision-making, and improve health outcomes.

Some researchers initially hoped that portals could reduce health inequities, defined as avoidable differences in health outcomes in vulnerable populations. Health inequities lead to poor health management and outcomes, which contributes to rising healthcare costs. Vulnerable populations often demonstrate lower health literacy and experience significant barriers to care, such as inflexible job hours, cost, and insurance status. Portal features might increase convenience, improve health literacy, and overcome at least some barriers to care, thereby reducing health inequities.

Unfortunately, more than 100 studies now show substantial health-equity-relevant disparities in portal use (citations available on request). Vulnerable populations use portals less often, including elderly persons, racial minorities, and persons with low socioeconomic status, low health literacy, chronic illness, or disabilities. Relatively low portal use in vulnerable populations may lead to intervention-generated inequity, a phenomena where well-intentioned solutions worsen existing health inequities rather than reduce them. Developing, implementing, and evaluating strategies to reduce disparities in portal use is critical to ensure portals benefit all populations as originally intended.

In this systematic review, we explore how researchers have confronted differential use of patient portals. Our review focuses on two critical questions: (1) what interventions impact portal use or predictors of portal use in vulnerable populations? (2) what interventions impact disparities in portal use?

Methods

We conducted this systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). A librarian searched Ovid MEDLINE, EMBASE, CINAHL, and Cochrane Reviews for eligible English-language research articles published before September 1st, 2018. Additionally, we searched tables-of-contents of pertinent scientific journals before December 1st, 2018 to identify recently published citations. We used Covidence systematic review software for citation screening, as recommended by Cochrane. Two reviewers independently selected English-language research articles that evaluated any interventions designed to impact an eligible outcome. One reviewer extracted data from the included studies, and another assessed accuracy. To assess risk of bias, we used predefined criteria from the AHRQ² to rate studies as low, medium, high, or unclear risk of bias. Two reviewers independently assessed risk of bias.

We conducted descriptive analysis of study characteristics and assessed intensity of intervention as per the Cochrane Handbook for Systematic Reviews of Interventions.³ Additionally, we categorized interventions according to the components described in the System Engineering Initiative for Patient Safety (SEIPS) model.⁴ The SEIPS model segments work systems into 5 tightly coupled components. Per the model, a person (component 1) performs a range of tasks (component 2) using various tools and technologies (component 3). Performance of tasks occurs within a physical environment (component 4) under specific organization conditions (component 5). Interventions may target one or more of the 5 components. We categorized interventions based on which component(s) were addressed. One team member categorized the interventions, and a second team member with experience applying the SEIPS model (NCB) reviewed the categorizations. Using the SEIPS model allowed us to determine gaps in the targets of current interventions and shortcomings related to failing to consider the interaction among components of the work system.
Results

Literature searches identified 719 potentially relevant citations. Of those, 91 studies were deemed eligible for full text review, and 18 studies fulfilled the inclusion criteria for this systematic review. Studies employed a broad variety of outcome measures over varied time periods, limiting their comparability. For example, when reporting portal use, studies variably reported login-days, total logins, activation, or another measure. Four out of 18 studies (22%) did not report on participants’ race, and 8 (44%) did not report on ethnicity.

Intensity of intervention varied widely across studies. An example low-intensity intervention was one-time assistance with credentialing, whereas an example high-intensity intervention was training participants across 4 weekly 2-hour sessions. Four of 18 studies (22%) had low risk of bias, 9 studies (50%) had medium, 3 studies (17%) had high, and 2 studies (11%) were unclear.

Figure 1 presents the SEIPS system components intervened on in each study. 13 of 18 studies (72%) intervened on the individual (person) component, 5 (28%) on the tool component (i.e. patient portal), 1 (6%) on the task component (e.g. prescribing portal content), 2 (11%) on the environment component, and 4 (22%) on the organization component. Seven studies (39%) intervened on 2 components, but no study intervened on more than 2.

Out of 18 included studies, 15 (83%) assessed an intervention’s impact on portal use and 7 (39%) assessed impact on predictors of use. Only one study reported on how an intervention impacted health-equity-related disparities in portal use. Ten out of the 18 studies (56%) were reports on how technical training or assistance for patients impacted portal use. Eight of the 10 reported or permitted calculation of statistical significance, of which 6 demonstrated benefit (i.e. the intervention increased portal use), 1 demonstrated neutrality (i.e. the intervention did not impact portal use), and 1 demonstrated mixed results (i.e. the intervention both increased and decreased aspects of portal use).

Discussion

A growing body of literature suggests that patient portals can prevent medical errors, increase shared decision-making, and improve at least certain health outcomes. Unfortunately, every year, studies continue to be published adding to the evidence of disparities in portal use, and interventions will be critical to ensure portals do not disproportionately benefit more advantaged populations. Our results suggest that few studies have evaluated interventions to reduce disparities in portal use. Due to the strong evidence for disparities in use, the limited research on addressing them, and the need to ensure all populations benefit from portals, we recommend that researchers shift from identifying disparities in portal use, 1 demonstrated neutrality (i.e. the intervention did not impact portal use), and 1 demonstrated mixed results (i.e. the intervention both increased and decreased aspects of portal use).

Out of 18 included studies, 15 (83%) assessed an intervention’s impact on portal use and 7 (39%) assessed impact on predictors of use. Only one study reported on how an intervention impacted health-equity-related disparities in portal use. Ten out of the 18 studies (56%) were reports on how technical training or assistance for patients impacted portal use. Eight of the 10 reported or permitted calculation of statistical significance, of which 6 demonstrated benefit (i.e. the intervention increased portal use), 1 demonstrated neutrality (i.e. the intervention did not impact portal use), and 1 demonstrated mixed results (i.e. the intervention both increased and decreased aspects of portal use).

Discussion

A growing body of literature suggests that patient portals can prevent medical errors, increase shared decision-making, and improve at least certain health outcomes. Unfortunately, every year, studies continue to be published adding to the evidence of disparities in portal use, and interventions will be critical to ensure portals do not disproportionately benefit more advantaged populations. Our results suggest that few studies have evaluated interventions to reduce disparities in portal use. Due to the strong evidence for disparities in use, the limited research on addressing them, and the need to ensure all populations benefit from portals, we recommend that researchers shift from identifying disparities in portal use to systematically addressing them. Additionally, we recommend that future studies measure interventions’ impact on disparities in use directly, as most studies to-date have not. Finally, categorization using the SEIPS model demonstrated that most interventions to-date addressed primarily the individual (person) component, and lacked coverage of the other components as well as combinations of components. To enhance impact, we recommend that future interventions affect, or at least consider the repercussions on, multiple components.

Upcoming Publication

This work was accepted at the Journal of the American Medical Informatics Association in February 2019. It is scheduled to be published in the November 2019 journal issue.

References

Relationship between adherence to daily home telehealth use and health services use in Veterans with heart failure

J Guzman\(^1,2,3\), B Viernes\(^4,5\), M Lee\(^1\), M Farmer\(^1\), B Wakefield\(^6,7\), M Yefimova\(^8\), & T Hahn\(^1,9,10\)
\(^1\)VA Greater Los Angeles HSR&D Center for the Study of Healthcare Innovation, Implementation & Policy; \(^2\)Southern Arizona VA Healthcare System, Tucson AZ; \(^3\)UCLA School of Nursing; \(^4\)VA VINCI, Salt Lake City UT; \(^5\)University of Utah; \(^6\)VA CADRE; \(^7\)University of Missouri; \(^8\)UCLA/VA HSR&D Clinician Scholar; \(^9\)UCLA School of Medicine; \(^10\)VA GLAHS GRECC

What will the attendee be able to do after being in your session?
At the end of this presentation, attendees will be able to identify potential predictors of emergency room (ER) visits and hospitalizations in Veterans enrolled in a home telehealth program (HT) for heart failure.

Description of the Problem or Gap: Use of HT for chronic disease monitoring and exacerbation management in the Department of Veterans Affairs (VA) is increasing, but little is known about the relationship between adherence to the daily use of HT devices and health service use outcomes, such as ER visits and hospitalizations.

Methods: Adherence to the use of HT by Veterans with heart failure who were enrolled in the VA Home Telehealth Program from 1/1/14 to 6/30/14 was obtained from VA Telehealth Services. Demographic, health service use, and facility-level variables were obtained from the VA Corporate Data Warehouse. Average adherence after 1 year of first device use was calculated from weekly reports of adherence (number of daily responses / expected responses or 7 days). Zero-inflated negative binomial (ZINB) regression was used to determine which variables influence the probability of having an outcome (ER visit; hospital admission) after HT enrollment and the expected frequency of the outcome occurring.

Results: A total of 3,449 Veterans met the inclusion criteria for this study. The majority were white (75%) and male (98%). Twenty-nine percent had a diagnosis of depression, 31% had a cardiac implant, and the average ejection fraction was 42.7% (SD=15.8). The average age was 71 years (standard deviation [SD] of 10.4) and the average Charlson Comorbidity Index (CCI) was 6.5 (SD=3.2). Almost 60% had a Care Assessment Need (CAN) score of 95% (high probability of hospital admission or death within 90 days at the time of HT enrollment). HT program size was divided in thirds with the largest HT program having between 850 and 1,495 patients. Average adherence was 57.1 (SD=30.6). There were less ER visits after HT enrollment - mean (SD) of 1.85 (2.8) the year after compared to the year before with 2.2 (3.4); 42% of the patients had no ER visits compared to the year before (33%). In addition, there were fewer hospital admissions after HT enrollment – mean (SD) of 1 (1.5) the year after compared to the year before with 1.2 (1.6); 56% of the patients had no hospital admissions the year after compared to the year before (44%).

Table 1. ZINB Model estimates – predicts probability of having a non-occurrence (dark cells were not significant in the bivariate analysis and thus were not included in the model for the specific health service outcome)*

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Emergency room visits</th>
<th>Hospital admissions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimates**</td>
<td>Exp (estimates)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-5.36 (NS)</td>
<td>0.005</td>
</tr>
<tr>
<td>Age</td>
<td>0.03</td>
<td>1.03</td>
</tr>
<tr>
<td>Race (non-White versus Black)</td>
<td>1.38</td>
<td>3.98</td>
</tr>
<tr>
<td>% Ejection Fraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large HT Program</td>
<td>0.83</td>
<td>2.30</td>
</tr>
</tbody>
</table>

Tables 1 and 2 shows the model estimates.
* Model also included (not significant): White vs Black, gender, CAN score, depression, comorbidity index, presence of cardiac implant, having prior ER visit or hospital admission before HT enrollment, rurality, average adherence and number of HT daily use after 1 year of enrollment; the hospital admission model also included reason for enrollment (e.g., non-institutional care).

** p<0.05; NS = not significant

Table 2. Count Model estimates – predicts the expected frequency of an outcome occurring (dark cells were not significant in the bivariate analysis and thus were not included in the model for the specific health service outcome)*

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Emergency room visits</th>
<th>Hospital admissions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimates**</td>
<td>Exp (estimates)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.21 (NS)</td>
<td>0.81</td>
</tr>
<tr>
<td>Race (White versus Black)</td>
<td>-0.15</td>
<td>0.86</td>
</tr>
<tr>
<td>CAN score &gt; 95%</td>
<td>0.48</td>
<td>1.62</td>
</tr>
<tr>
<td>Has depression</td>
<td>0.13</td>
<td>1.13</td>
</tr>
<tr>
<td>% Ejection Fraction</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Moderately high CCI</td>
<td>0.27</td>
<td>1.31</td>
</tr>
<tr>
<td>High CCI</td>
<td>0.36</td>
<td>1.43</td>
</tr>
<tr>
<td>Has cardiac implant</td>
<td>0.12</td>
<td>1.13</td>
</tr>
<tr>
<td>Had 1+ ER visit before HT use</td>
<td>0.63</td>
<td>1.89</td>
</tr>
<tr>
<td>Had 1+ hospital admission before HT use</td>
<td>0.32</td>
<td>1.37</td>
</tr>
<tr>
<td>Number of HT reports submitted</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

* Model also included (not significant): age, white versus other minority (non-black), gender, program size, rurality, average adherence after 1 year of enrollment; the hospital admission model also included reason for enrollment (e.g., non-institutional care).

Discussion: Patient adherence to the use of home telehealth devices is important to the success of any home telehealth programs to support chronic disease monitoring and management. In this study, average adherence after 1 year was not associated with ER visits or hospital admissions. Older age, being from a minority group (non-Black), and belonging to a large HT program were associated with not having an ER visit a year after HT enrollment. Being in poorer health (i.e., high CAN score, having depression, high CCI, having prior ER visit, etc.) was associated with higher expected count of going to the ER and having a hospital admission holding all other variables constant after HT enrollment. Limitations of this study include those inherent in the use of secondary data, use of VA data only, and inclusion of only those Veterans with heart failure enrolled in the first part of 2014.

Conclusion: Adherence to the use of HT devices was not found predictive of ER visits or hospital admissions in the year after HT enrollment. Other outcomes, such as quality of life, patient satisfaction, and off-load of work from the primary team, may be better outcomes to consider success of HT programs. Depression in patients with heart failure enrolled in HT programs should be addressed as this is associated with health service use. Further research is needed to identify factors related to race differences in health service use, as well as organizational factors that affect health service use in Veterans enrolled in HT programs.
Design and Use of an Inpatient Peer Support Technology Probe

Shefali Haldar¹, Yoojung Kim¹², Ari H Pollack¹³, Wanda Pratt¹
¹University of Washington, Seattle, WA; ²Seoul National University, Seoul, South Korea; ³Seattle Children’s Hospital, Seattle, WA

Introduction
When patients are admitted to the hospital, they have many needs—such as understanding their care, and communicating with their providers—that manifest over the course of their stay. Some technologies have helped to address these care-specific needs,¹ but inpatients have additional informational, emotional, and experiential needs that are not currently fulfilled. Patient-peers can serve as a critical resource to meet these additional needs.² Peer support—the exchange of advice among patients with common health experiences—is attributed to many positive outcomes, such as increased self-efficacy, knowledge, and empowerment.³ Prior work has demonstrated a strong desire to exchange peer support among inpatients.⁴⁻⁵ Yet, technologies that facilitate peer support (e.g., online health communities) are commonly studied in every-day life and outpatient settings, or with patients managing chronic conditions. Although peer support programs have been created for both pediatric and adult inpatients, most involve in-person or phone encounters, and the potential for a technology-driven approach has not yet been fully realized. Therefore, we sought to investigate how patients use a peer support tool to interact with each other in the real-world context of their hospitalization.

Methods
We used the technology probe method, a Human-Computer Interaction approach to deploy a technology (i.e., probe) in a real-world context, study how it is used, and examine its impact on end-users.⁶ We first designed and built a fully functional inpatient peer support technology probe, then deployed it at a pediatric and an adult hospital. Patients were eligible if they were at least 7 years old, well enough to provide informed consent, and could communicate in English. Parental consent was obtained for all pediatric patients. We enrolled 30 participants (15 at each site) who were evenly divided among gender identity (female and male) and hospital service (medical and surgical). At the pediatric site, the mean age was 13 years (SD 2.56) and mean length of stay was 5.13 days (SD 3.74). At the adult site, the mean age was 55.8 (SD 15.84) and mean length of stay was 14.4 days (SD 22.79). We collected data about participants’ usage period of the probe, lasting from the time of their enrollment in the study until their discharge from the hospital. All study procedures were approved by our IRB.

Inpatient Peer Support Technology Probe Design
By using a subset of design recommendations we previously identified on this topic,⁴⁻⁵ we created an inpatient peer support technology probe that allowed an asynchronous exchange of peer advice among inpatients. We designed a web-based probe resembling an online health community with functional similarities to Reddit, a web content repository and social media site. Key features included stories (original posts), comments, and votes (analogous to “likes” on other social media sites). Our probe was built using WordPress and hosted at two domain names, one designated for each study site. Links to the probe were only given to study participants. All participants were given access to the probe via a research team-owned iPad.

Analysis of Usage Period Data
Descriptive statistics were used to analyze data about participants’ usage events on the probe (i.e., the number of stories, comments, and votes they contributed to the system). Using our previously identified peer support needs of inpatients,⁴⁻⁵ we did a deductive qualitative analysis of the content they contributed to the probe (i.e., the type of information shared with peers through posting stories and comments). To measure the emotional tone of these stories and comments, we conducted a sentiment analysis using the IBM Watson Natural Language Understanding (NLU) tool. Finally, we did a Social Network Analysis (SNA) to analyze the interactions that participants had with each other (i.e., comments in response to stories; which participants voted on which stories).

Results
Throughout our study period, 76% (23/30 participants) contributed to the technology probe. Pediatric participants averaged a total of 2.7 usage events and posted 13 stories, 2 comments, and 12 votes. Their adult counterparts averaged
a total of 2.2 usage events and contributed 6 stories, 14 comments, and 13 votes. Across both study sites, a total of 6 participants exclusively posted stories, 1 participant exclusively commented on stories, 5 participants exclusively voted on stories, and 11 participants did some combination of those three types of activities. The remaining users (2 pediatric and 5 adult participants) had no recorded usage events on the probe due to illness or disinterest in the content. Participants shared a variety of information and support for their peers, including how to adjust to the unfamiliar hospital environment (“if you are staying the night in the hospital, pack your own comfy blanket”) and encouragement to speak up to providers about a problem (“if you don’t think you’re getting the care you need, ask for the charge nurse as recourse”). Stories and comments of 81.8% of pediatric participants and 57.1% of adult participants received positive mean sentiment scores. Our SNA revealed pediatric participants’ interactions were relatively spread out and contained 80% (12/15) participants, while adult interactions contained 60% (9/15) participants, gravitating towards 2 primary interaction “hubs”.

Discussion and Conclusion

Our participants’ use of the probe suggest that technologies designed to enable inpatient peer support have great potential in this context. Our analysis revealed that pediatric and adult participants share advice that is largely encouraging to peers and positive in emotional tone. Moreover, each site had different usage patterns: pediatric patients shared individual stories and had distributed interactions, while adult patients had concentrated interactions with existing discussions. These distinct usage patterns might be due to generational differences regarding the perceived purpose and value of technologies resembling social media. Future inpatient-facing technologies must be designed to accommodate these differing usage patterns, as technologies must support broader patient populations. In addition, future work should explore strategies for long-term implementation of peer support technologies within hospitals.

References

Simulation Study of Just-in-Time Specimen Recruitment from University-Wide e-Phenotypes of Interest

Bashir Hamidi, MPH, Leslie A. Lenert, MD, FACMI, Alexander V. Alekseyenko, PhD
Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC

Introduction
The ability to obtain for human specimens from precisely phenotyped cohorts is important for advancing translational research. A pool of suitable specimens, including blood and microbiology, are often available within medical centers incidental to delivery of routine care. After recording of the clinical test results, surplus aliquots of the biospecimens are usually available to be re-routed for research, albeit for a short window of time before they are discarded. Biobanking of such specimens could preserve them for future research use; however, the burden with setting up and maintaining blood or microbiology biobanks may be too large for many institutions. Therefore, systems and platforms for just-in-time re-use of surplus clinical specimens are necessary to minimize the regulatory burden for investigators and accelerate translational research. These systems and processes should allow investigators to perform population and cohort-based studies with de-identified specimens that qualify as non-human subjects research (exemption 4). Examples of existing platforms include: (i) Harvard’s Crimson system, which combines its i2b2 instances with a link to its laboratory specimen processing system; and (ii) the Vanderbilt BioVu system, which operates more like a traditional biobank with storage of specimens.

In conjunction with the MUSC Living BioBank (LµBB) system for surplus blood, we are developing the Living µBiome Bank (LµBB), a NIH/NCATS supported system for just-in-time capture of existing microbiology specimens for microbiome research. Both systems, LµBB, operate as a unit and rely on the state-of-the-art concepts in (i) e-phenotyping, to identify cohorts of patients for which specimens are necessary; (ii) research permissions capture, to ensure ethical and compliant specimen recruitment; and (iii) research data warehousing (RDW), to store de-identified microbiome measurements linkable to rich clinical data. The primary microbiology specimen source for the LµBB is the active infection surveillance culture program for methicillin-resistant Staphylococcus aureus and vancomycin-resistant Enterococcus. Annually, this program generates over 15,000 nasal swabs, which can serve as proxy for the oral and upper-respiratory microbiome, and 7,500 peri-anal swabs, which can serve as practical proxy for gastrointestinal microbiome. These swabs are collected incident to treatment and typically early in the treatment course. The surveillance testing typically results in 75% of the specimen volume remaining as surplus, which are suitable for marker gene sequencing-based microbiome assays.

Methods
The adoption of just-in-time biobanking concept relies on the ability of an investigator to specify electronic phenotypes (e-phenotypes) correctly and ensure that the resulting cohort captures the nuanced clinical characteristics of the desired study population. We have executed an expert-driven simulation study of the LµBB to understand the challenges involved in specification of e-phenotypes and to develop methods to estimate phenotype-specific accrual rates by simulating recruitment of specimens from retrospective RDW data. Many routinely collected specimens can be included in LµBB, but here we have focused on three specimen types: (i) blood; (ii) nasal swabs; and (iii) peri-anal swabs. MUSC IRB considered this a quality improvement project and therefore exempt from review.

Figure 1. Simulation study outline (left) and data obtained from RDW (right).
The outline of the simulation study design is shown in Figure 1A. We have identified 20 experts across MUSC, including College of Medicine, College of Dental Medicine, College of Nursing, and Cancer Center to serve as disease area experts to conduct the simulation study of 20 respective e-phenotypes and cohorts. The experts have completed an entry questionnaire to evaluate their degree of proficiency with i2b2 used for e-phenotyping. With assistance from study coordinator and honest broker each expert has specified their desired e-phenotype in i2b2 and transmitted it to Biomedical Informatics Center (BMIC), where the query has been translated into SQL query into RDW data. BMIC database team has simulated recruitment of 20 subjects matching the e-phenotype in historical de-identified RDW data. We have produced de-identified charts for these patients, along with accrual rates for the specimens and other relevant metrics (Figure 1B). Each chart has been reviewed by the expert to determine if it is a match or mismatch to the desired phenotype, or if it is hard to tell. An exit questionnaire has been offered to the expert to reflect on their experience using i2b2 system and their recommendations for the specimens obtained from L(µ)BB.

Results

The entry questionnaire included a single question aimed at gauging the experts' level of confidence in being able to specify an e-phenotype of interest using i2b2. Using a four-level Likert scale only 20% of experts were not at all confident, 15% were highly confident. Based on their feedback, most of the respondents conflated their own ability to use the i2b2 system, with expectations about the quality of the data, and their confidence in study coordinator and honest broker, which could have possibly confounded these results. The data simulated from RDW reproduced the total number of observed patients from i2b2 with correlation of close to 100%. The largest relative discrepancy in the numbers was observed in phenotypes with the least number cases.

For phenotypes where the experts have been able to review charts, the chart match rate to the desired phenotype varied dramatically (Fig. 2A). Interestingly, the mean match rate was inversely proportional to the expert confidence level (Fig. 2B). Most of the e-phenotypes yielded cohorts where blood and microbiome specimen accrual rates are reasonable for a pilot study with quick turnaround (Fig. 2C).

Discussion

This project revealed several issues with defining e-phenotypes for just-in-time biobanking: (i) e-phenotypes descriptions need to differentiate patient characteristics vs clinical events necessary for patient inclusion/exclusion; (ii) temporal constraints are hard to specify and are often not intuitive even for experienced i2b2 users; (iii) many data issues may affect the resulting phenotype and cause the expert to compromise their desired definition bases on the mapping status of diagnoses, problem lists, labs and heterogeneity of coding procedures by different clinical teams; (iv) medications are perhaps the most challenging single group of phenotype characteristic, especially because the route of administration often is not documented. In our talk, we will discuss these, and other issues identified in the project and present more details on the results. The details of IRB status of the L(µ)BB will likewise be discussed.

References

Effects of an Interactive Trust-enhanced Electronic Consent on Patient Experiences with Consenting to Share their Health Records for Research

Christopher A. Harle, PhD1,2, Elizabeth H. Golembiewski, MPH1, Kiarash P. Rahmanian MPH3, Babette Brumback, PhD4, Janice L. Krieger, PhD5, Kenneth W. Goodman, PhD6, Arch G. Mainous III, PhD3,7, Ray E. Moseley, PhD3

1Department of Health Policy and Management, Indiana University, Indianapolis, IN; 2Regenstrief Institute, Indianapolis, IN; 3Department of Community Health and Family Medicine, University of Florida, Gainesville, FL; 4Department of Biostatistics, University of Florida, Gainesville, FL; 5Department of Advertising, University of Florida, Gainesville, FL; 6Miller School of Medicine, University of Miami, Miami, FL; 7Department of Health Services Research, Management and Policy, University of Florida, Gainesville, FL

Introduction

Electronic health record (EHR) data has created many opportunities for research, including analysis of large-scale repositories, targeted research recruitment, and cross-institutional data sharing.1,2 However, ethical and legal standards suggest and sometimes mandate that patients consent to the use of their health information for secondary research purposes. For example, the U.S. Federal Policy for the Protection of Human Subjects (Common Rule) now permits health care institutions to seek broad consent from patients for future, to-be-specified, research uses of their EHR data.3 Broad consent has potential value for institutions, as it may facilitate the use of patient data to analyze treatment and outcomes over time, and it gives patients more control over their health records compared to waivers of informed consent. However, little guidance exists for best practices for obtaining broad consent in a way that is minimally invasive to provider workflows and easily scalable, while also ensuring that patients have the information needed to make an informed decision. In addition, little is known about how to present consent information electronically (i.e., as an e-consent) in order to maximize patients’ satisfaction with their consent decisions and their understanding of consent content. Therefore, in the context of broad consent for future research using identifiable EHR data, the objective of this study was to compare the effectiveness of an interactive trust-enhanced broad e-consent to an interactive only e-consent and a standard e-consent (no interactivity, no trust enhancement).

Methods

A randomized trial was conducted among adult, English-speaking participants making a regularly scheduled primary care visit. Each participant was given a tablet computer and block randomized to one of three e-consent application conditions, which were previously developed via a user-centered design process.4 Each condition asked patients for their consent to include their identifiable EHR data in a family medicine database for use in future, yet-to-be-specified research studies. The standard e-consent contained only the minimum federally-required elements of informed consent. The interactive-only e-consent included all information present in the standard e-consent, as well as clickable links to additional content. Finally, the interactive trust-enhanced e-consent included all elements of the standard and interactive only e-consents, as well as five short messages describing trust-relevant information on research regulations, researcher training, and data protections. Primary outcomes were patient-reported satisfaction (modified Satisfaction with Decision scale) with and subjective understanding of the e-consent (modified Quality of Informed Consent scale). Secondary outcomes were objective knowledge, perceived voluntariness, trust in medical researchers, and consent decision. All measures were based on existing scales found in peer-reviewed literature.5 Outcomes were assessed immediately after e-consent use, and at one-week and six-month follow-up. Mixed model regression analyses were used to estimate the relationship between consent condition and each outcome over time. Each model included a participant random effect and fixed effects for consent condition, time (immediate, one-week, or six-month follow-up), participant age, education level, gender, race, and ethnicity.

Results

At immediate follow-up, participants across all conditions (N = 734) reported moderate to high satisfaction with consent (mean 4.3 out of 5) and subjective understanding (79.1 out of 100; see Figure 1). In addition, over 94% of
participants consented to share their health record data. At immediate and one-week follow-up, no differences in primary or secondary outcomes were observed between conditions. However, at six-month follow-up, relative to participants who used the standard e-consent, participants who used the interactive, trust-enhanced e-consent reported higher satisfaction ($\beta = 0.86, p < 0.0001$) and subjective understanding ($\beta = 31.94, p < 0.0001$) as well as higher objective knowledge ($\beta = 2.11, p < 0.0001$), perceived voluntariness ($\beta = 3.87, p < 0.0001$), and trust in medical researchers ($\beta = 12.38, p < 0.0001$). Similarly, at six-month follow-up, participants who used the interactive only version reported higher satisfaction, subjective understanding, objective knowledge, and trust in medical researchers relative to participants who used the standard e-consent (all $p < 0.0001$). In sensitivity analyses to account for attrition, similar results were obtained when examining only the subset of participants (N = 472, 64%) who participated in all three follow-up surveys.

Figure 1. Satisfaction (1-5) and subjective understanding (0-100) at immediate, one-week, and six-month follow-up.

Discussion

In a randomized trial of three e-consent applications, a large majority of participants agreed to share their health information for future unspecified research. At six-month follow-up, patients who had used an e-consent with only standard information exhibited reductions in their initially high satisfaction with and understanding of the consent. Conversely, participants who had used an e-consent with interactive capabilities to obtain additional information and with messages designed to enhance trust in research institutions and processes tended to maintain high levels of satisfaction, understanding, knowledge, perceptions of voluntariness, and trust. These results suggest that systems designers, policymakers, and institutions that implement broad consent may achieve better patient consent experience outcomes by augmenting standard consent content with options to obtain more information and with factual messages designed to enhance trust by addressing patient concerns about how their data are used and safeguarded in research. As patient-facing informatics applications continue to integrate with healthcare delivery processes and information technology, e-consent tools are increasingly feasible. Still, the value of implementing and maintaining e-consent for research use of health records must be weighed against other, potentially conflicting operational objectives, such as patient throughput.

References

Is Online Portal Training Effective for Older Adults? An Experimental Study to Evaluate the Impact of Providing Patient Portal Multimedia Instructional Materials on Older Adults' Performance in Using Portals

Yuqi He, MLIS
University of Wisconsin - Madison, Madison, WI

Introduction

Patient portals (PPs) are recognized as one of the most important tools to solve the challenges of accessing and managing personal health information and support greater patient engagement. The growing aging population could greatly benefit from using PPs because of the prevalence of chronic conditions and high medication use, and therefore a more pressing need for ongoing health information management and self-management of conditions. However, when compared with other age groups, older adults are generally lagging behind others on portal enrollment and portal use. It is likely that older adults encounter more problems in using portals for a number of reasons, including age-related sensory and cognitive declines, increased risk of functional impairments, less experience with computers and internet, and higher prevalence of low health literacy and numeracy1.

One critical barrier to PP adoption is lack of sufficient training to instruct patients on use of a portal2. Research has shown that older adults with limited computer experience and limited health literacy3 require additional support to learn how to use the system. Studies also found that even older adults with adequate health literacy and internet experience desire instructions on specific tasks for using portals3. In spite of a strong need for portal training, this training is generally not given by health care providers due to lack of resources4.

In an effort to fill the gap on PP instructional support, we designed two forms of online PP multimedia instructional materials for new older portal users (see 5 for the design process). One was a PDF file using a combination of graphics (i.e., screenshots) and texts; another was a customized slide-show presentation (in HTML5 format) using graphics and narrations together. The aim of this study was to evaluate their outcomes to explore the value of training. In particular, we hypothesized that 1) participants who received training would demonstrate better outcomes than the control group in PP knowledge, PP self-efficacy, and usability metrics (task success, task time, task difficulty, and perceived usability), and 2) training would result in best performance on tasks identical to tasks learned in the instructional materials and worst performance on tasks different from tasks learned in the instructional materials.

Methods

Experimental Design: A randomized complete block design (RCB) was used to reduce variability within treatment conditions. Health literacy and computer experience were chosen to serve as blocking factor because these two factors have been identified as critical variables that affect portal use among older adults1,3. The experiment had three independent variables: block with four levels (inadequate health literacy (-HL) & inadequate computer experience (-CE), -HL & +CE, +HL & -CE, +HL & +CE), treatment/instruction with three levels (HTML5 instruction, PDF instruction, no instruction/control), and task type with three levels based on the closeness between the task problems and the tasks learned in the instructional materials (identical tasks, similar tasks, different tasks).

Eligibility: All participants were required to 1) be aged 65 years or older, 2) be native English speakers, 3) do not have a diagnosis of cognitive impairment, 4) be able to read a computer screen, 5) be able to hear sound from a computer, 6) be able to use a keyboard and a mouse, (7) and have not used a patient portal before (novice users).

Procedure: The researcher traveled to community places at participants’ choice for data collection. At the beginning of a session, participants were asked to complete a battery of questionnaires on demographics, computer and internet experience, health literacy6, and baseline PP knowledge. Next, the researcher provided basic computer training to the participants on the study laptop. Then, Participants were randomly assigned to one of the three treatment/instruction conditions. Upon the completion of instructions, participants were asked to complete 15 tasks (5 identical, 5 similar, and 5 different tasks) on a portal and the Single Ease Question (SEQ)10 appeared after each task for measuring perceived difficulty on a task level. Task time and participants’ on-screen activities were captured by a usability testing software (i.e., Morae) for later review. Last, participants completed questionnaires on System Usability Scale (SUS)11, PP knowledge, and PP self-efficacy (modified based on12). The whole research session lasted 90 minutes on average. The portal used for this study was a simulated patient portal based on Epic's MyChart, named "MyHealth Portal."
Results

Forty participants completed this study (the required sample size from power analysis was 36), including 14 men (35%) and 26 women (65%). The majority of them (80%) were White. Participants’ age ranged from 65 to 86 years ($M = 73.07, SD = 6.61$). The vast majority of them (95%) were retired and reported to be in good to excellent health. The sample was educationally diverse and had fairly low income.

Figure 1 below shows the percentage of participants (in trained and not trained group) who completed each task successfully. In both groups, the most successful task was Task 1 (identical task) and Task 5 (similar task). In contrast, the least successful task was Task 4 (different task) and Task 9 (different task). Analysis on task time and task difficulty revealed that Task 4 and Task 9 took the longest time and were rated as the most difficult tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>log in</td>
</tr>
<tr>
<td>2</td>
<td>find out a health issue</td>
</tr>
<tr>
<td>3</td>
<td>locate a link to educational materials</td>
</tr>
<tr>
<td>4</td>
<td>find out the last blood pressure reading</td>
</tr>
<tr>
<td>5</td>
<td>locate the medication page</td>
</tr>
<tr>
<td>6</td>
<td>find out Lipitor dosing instructions</td>
</tr>
<tr>
<td>7</td>
<td>locate the test results page</td>
</tr>
<tr>
<td>8</td>
<td>check HbA1c in a standard range</td>
</tr>
<tr>
<td>9</td>
<td>find out the graph of sodium level change</td>
</tr>
<tr>
<td>10</td>
<td>check osteoporosis screening</td>
</tr>
<tr>
<td>11</td>
<td>schedule a visit</td>
</tr>
<tr>
<td>12</td>
<td>cancel an appointment</td>
</tr>
<tr>
<td>13</td>
<td>send a message</td>
</tr>
<tr>
<td>14</td>
<td>view all sent messages</td>
</tr>
<tr>
<td>15</td>
<td>view a medical bill</td>
</tr>
</tbody>
</table>

Figure 1. Success rates for 15 tasks

Statistical analysis showed that, when controlled for health literacy and computer experience, older participants who received portal training had significantly better outcomes in after-test PP knowledge scores ($p < .001$), PP self-efficacy ($p < .001$), total task success ($p < .001$), and SUS ratings ($p = .003$) than those who had no training, except for total task time ($p = .37$) and total SEQ scores ($p = .22$). The HTML5 instruction did not result in any significantly better outcomes than the PDF instruction ($p > .05$). These results indicate that the HTML5 instruction and the PDF instruction were equally effective in improving older adults’ performance in using patient portals.

Mixed ANOVAs were constructed to examine group (trained and not trained)-by-task type interaction for testing H2. The interaction term was significant on task success ($p = .03$) and SEQ scores ($p = .004$), but not task time ($p = .79$). These results suggest that, depending on the instructional content, training can result in best task performance (more tasks completed, lower task difficulty) on identical tasks; however, training has no notable effect on tasks that are different from the tasks learned in the instructional materials.

Conclusion

The findings of this study show that PP instructional materials are effective in improving older adults’ knowledge and confidence in using PP, and boosting task success rate and perceived usability of a PP website.

References


1362
Redesigning PopMedNet™ for distributed regression analysis with vertically partitioned data

Qoua L. Her, PharmD, MSc¹, Yuji Samizo, MS², Thomas Kent, PhD², Aleksandra Slavkovic, PhD², Mia Gallagher, MPH¹, Sengwee Toh, ScD¹

¹Harvard Pilgrim Health Care Institute, Boston, MA; ²Department of Statistics, Penn State University, University Park, PA

Introduction

Generally, multiparty analyses involve two types of data partition environments, horizontal and vertical. Horizontally partitioned data refer to an environment where data-contributing sites hold distinct cohorts of patients with the same data attributes (Figure 1).¹ Vertically partitioned data refer to an environment where data-contributing sites hold distinct data attributes of a cohort of patients (Figure 2). Distributed regression analysis (DRA) is a privacy-protecting analytical method that allows investigators to perform multivariable regression analysis in both types of data environments, with only summary-level data.¹ DRA has been demonstrated to produce precise regression parameter and standard error estimates (<10⁻⁶) compared to the estimates obtained from regression analysis with the pooled patient-level data.¹,² However, DRA with some regression models requires an iterative process, where data-contributing sites compute and share summary-level data with a trusted-third party (often refer to as the analysis center) until the model converges.

This iterative process is laborious and susceptible to human error. We previously enhanced PopMedNet™, an open-source file-sharing software, to create an automatable and iterative file transfer process, to organize and facilitate DRA with horizontally partitioned data.³ We integrated the workflow with a SAS-based DRA application and perform DRA with horizontally partitioned data. Limited work has been done to facilitate DRA with vertically partitioned data. We explore the feasibility of using PopMedNet to organize and facilitate DRA with vertically partitioned data.

Methods

We gathered the statistical and informatic requirements of DRA with vertically partitioned data. We developed an R-based application to perform distributed linear, logistic, and Cox proportional hazards analysis with vertically partitioned data. We integrated the application with PopMedNet and identified and implemented additional PopMedNet enhancements needed to optimize the execution of the application.
**Results**

PopMedNet can be used to facilitate DRA with vertically partitioned data. Initially, we were able to compute precise regression results (< $10^{-7}$) with datasets of small sample sizes and few covariates (n < 5,000 and 10 covariates). DRA with larger datasets (more patients and covariates) would require significantly longer computational (> 24 hours) and file transfer times (> 24 hours), making PopMedNet inoperable with its existing configuration. To decrease the computational time of large datasets, we horizontally partitioned the two vertically partitioned datasets into blocks, performed the matrix computations within each block, and aggregated the computed summary statistics across the blocks. This approach greatly reduced the computational time, but at the tradeoff of increased overall file transfer time due to a greater number of files needed to be transfer, again making PopMedNet inoperable.

Thus, we implemented two new PopMedNet enhancements to optimize and facilitate the execution of the DRA application with vertical partitioned data. The first was concurrency of file upload and download. Prior to this enhancement, PopMedNet uploaded and downloaded only one file at a time. Concurrency increased the number of simultaneous file upload and download, which decreased overall file transfer times. The second was the option to transfer files directly between data-contributing sites. We previously designed PopMedNet to only transfers files between data-contributing sites and the analysis center. This enhancement allows PopMedNet to support DRA algorithms for vertical partitioned data that are computational more efficient. We were able to perform vertical DRA with datasets of 20,000 to 1,000,000 observations and five to 45 covariates, with precise regression results (< $10^{-7}$) compared to the pooled patient-level data analysis (Table 1). Full end-to-end execution times ranged from 6 to 92 minutes, depending on the regression model type, size of the dataset, and the number of covariates in the model.

**Discussion**

PopMedNet can be used to facilitate vertical DRA. To increase computational and operational performance, two new PopMedNet functionalities were developed and implemented.

**Table 1: Distributed Logistic Regression vs. Pooled Individual-level Logistic Regression (N = 1,023,061)**

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Distributed Regression Analysis</th>
<th>Pooled Individual-level Data Analysis</th>
<th>Difference in Parameter Estimates</th>
<th>Difference in Standard Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter Estimate</td>
<td>Standard Error</td>
<td>Parameter Estimate</td>
<td>Standard Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>-5.9550</td>
<td>0.0340</td>
<td>-5.9550</td>
<td>0.0340</td>
</tr>
<tr>
<td>Surgery</td>
<td>0.0070</td>
<td>0.0060</td>
<td>0.0070</td>
<td>0.0060</td>
</tr>
<tr>
<td>Female</td>
<td>0.0070</td>
<td>0.0080</td>
<td>0.0070</td>
<td>0.0080</td>
</tr>
<tr>
<td>Race</td>
<td>-0.0170</td>
<td>0.0110</td>
<td>-0.0170</td>
<td>0.0110</td>
</tr>
</tbody>
</table>

*Results of the 45 covariates are not shown, all DRA parameter and standard error estimates are precise to the estimates from the pooled patient-level data analysis (<$10^{-7}$)*

**References**


**Acknowledgements**

Research reported in this abstract was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number U01EB023683. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Prior to issuing a press release concerning the outcome of this research, please notify the NIH awarding IC in advance to allow for coordination.
Building Safer EHRs: Hospital Medication Order Safety Performance

A Jay Holmgren, MHI,1,2 Zoe Co, BS,2 Lisa P. Newmark, BA,3 Melissa Danforth, BA,4 David C. Classen, MD MSc,5 David W. Bates, MD MSc.2,3,6

1Harvard Business School, Boston, MA, 2Department of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, 3Clinical and Quality Analysis, Partners Healthcare, Somerville, MA, 4The Leapfrog Group, Washington D.C., 5University of Utah Health Science Center, Salt Lake City, UT, 6Harvard Medical School, Boston, MA

EHRs can improve safety via decision support-enabled CPOE, to catch adverse drug events (ADEs) before they reach the patient. A national sample of 1,527 hospitals from 2009-2016 took a safety performance assessment using simulated medication orders. The average EHR identified 54.0% of potential ADEs in 2009, rising to 61.6% in 2016. Hospitals which took the assessment multiple times performed better in subsequent years. The details of implementation and use of CPOE plays a key role in performance.

Introduction: Electronic health records (EHR) with decision support-enabled computerized physician order entry (CPOE) have become ubiquitous in US hospitals. CPOE allows decision support systems to intervene at the point of care, preventing potential adverse drug events (ADEs) before they reach the patient. However, evidence indicates that the effectiveness of these systems at improving quality is mixed.1 Hospitals have tremendous latitude with respect to what they actually implement with any vendor, especially around decision support. It is clear that EHR adoption does not necessarily lead to patient safety improvements – the structural quality improvement (EHR implementation) does not necessitate the process quality improvement (ensuring CPOE systems identify ADEs).2 We used a national hospital sample with data from a simulation designed to assess clinical decision support systems performance in identifying potential adverse drug events in hospital CPOE systems, to answer three research questions. First, how effective are hospital CPOE systems at identifying potential adverse drug events? Second, has hospital CPOE safety performance improved over time and as hospitals take the test repeatedly? Finally, what hospital characteristics are associated with better CPOE safety performance?

Methods: We used data from the Leapfrog Group’s CPOE assessment test from 1,527 hospitals in an unbalanced panel from 2009 to 2016 with 5,107 hospital-year observations. The Leapfrog CPOE assessment uses simulated patients and medication orders to evaluate what percentage of potential adverse drug events were correctly identified by an electronic health record.3 We linked these results with data from the American Hospital Association (AHA) Annual Survey from 2009 – 2016 to capture hospital demographic information. We first calculated mean CPOE assessment test scores over time, from 2009 through 2016. We then created a multivariate ordinary least squares regression model with hospital CPOE test scores as our dependent variable and our set of hospital demographics as independent variables, as well as number of years the hospital had participated in the CPOE simulation test. Our model also included hospital random effects and year fixed effects, as well as robust standard errors clustered at the hospital level.
Results: The mean CPOE performance test result increased over time, starting at 54.0% in 2009, and rising over time: 57.1% in 2010, 56.8% in 2011, 58.1% in 2012, 58.7% in 2013, 58.7% in 2014, 60.2% in 2015, and finally 61.6% in 2016. (Figure 1)

Figure 1. Mean Hospital CPOE Safety Performance Test Scores, 2009 – 2016

In the multivariate regression results, we found that each year of experience in taking the test was significantly associated with a 1.9% increase in CPOE safety performance scores ($p < 0.001$), even after controlling for other hospital demographics such as size, academic medical center status, location, and ownership.

Conclusion: While hospital performance on the Leapfrog CPOE assessment improved from 2009 to 2016, decision support systems still caught fewer than two-thirds of potential ADEs in simulated patient orders. Hospitals who took the assessment multiple times performed better in subsequent years than hospitals taking the test for the first time, suggesting that intentional quality improvement efforts as represented by participating in the voluntary CPOE assessment are an important enabler of higher safety performance.

References
An information infrastructure for federated, person-level linkage and query capability across private- and public-sector health data: The Indiana state EMS to HIE ED pilot project

Shaun J. Grannis, MD, MS1,2; Daniel R. Hood, MPH1; Katie S. Allen, BS3; Josh L. Martin, MPA3; John M. Roach, MS, MBA3,4; Darshan N. Shah, MBA3; Peter J. Embi, MD, MS1,2

1Regenstrief Institute, Indianapolis, IN; 2Indiana University School of Medicine, Indianapolis, IN; 3Indiana Management Performance Hub; 4KSM Consulting

Introduction

The era of precision health recognizes the increasing value of more diverse and granular data elements to inform our understanding of human health.1 In support of this approach, there is demand for new approaches for integrating clinical, social, genetic, environmental and other exposome data for individuals and populations. Regenstrief has partnered with several organizations addressing Indiana’s opioid epidemic, including the Indiana Management Performance Hub (MPH), to create both a technical infrastructure and a governance process to facilitate gathering, managing, sharing, and using population health data for both practice and research in Indiana. The technical infrastructure leverages a common data model that reconciles individual and population-level data. The governance model, which includes representation from stakeholders that both contribute and consume data, develops policies and processes for exchanging key data for priority use cases, including individual and population level data sets focused on opioid-related abuse and deaths.

The Regenstrief Institute curates many clinical data sources including electronic health information from over 100 institutions spanning Indiana.2 The Indiana MPH enables data sharing and transparency among myriad state organizations, maintaining data from the Department of Homeland Security, Indiana Commission for Higher Education, Department of Workforce Development, State Police and others, which can be leveraged in public-private collaborations to enable data-driven decision making. Regenstrief and MPH partnered to pilot a data integration project to enhance the usability of clinical and social determinants of health data to study the opioid epidemic in Indiana. This presentation describes the process for matching patient-level emergency department visits with Indiana Department of Homeland Security emergency medical service (EMS) run data.

Methods

EMS run data was made available for the years 2014 through 2017 by the Indiana Department of Homeland Security (DHS) and the Indiana MPH. From 2014 through 2017 a total of 2,228,796 EMS runs were reported through the Indiana DHS and attributed to 1,258,294 individuals. An emergency department utilization cohort was generated from the Indiana Network for Patient Care (INPC), one of the nation’s largest health information exchanges. Individuals with one or more emergency department (ED) encounters from 2014 through 2017 were included in the study cohort. From 2014 through 2017, a total of 18,204,788 ED encounters were reported through the INPC and attributed to 3,169,999 of individuals.

We used a privacy preserving record linkage (PPRL) process to protect unique patient identifiers within both cohorts.3 Both Regenstrief and MPH applied the hashing methodology to their respective cohorts. In addition to the hashed identifers, we created a unique pseudonymous identifier for each individual, which maps back to each source’s original data. Regenstrief securely transmitted hashed identifiers and pseudonymous ID’s for the INPC cohort to Indiana MPH. After identifying matching individuals with one or more EMS run, MPH then securely transmitted the EMS data payload to Regenstrief with the Regenstrief pseudonymous ID attached. (Figure 1)

![Figure 1: Data Matching Process. Regenstrief and MPH each created and exchanged local hashed identifiers, identified matching records, and created a composite payload dataset.](image)

Results

The matching process yielded a total of 244,421 unique individuals linked between the two data sets. Of these linked individuals, a total of 331,666 ED encounters could be linked to corresponding EMS runs. Table 1 highlights the yearly breakdown of the linked ED encounters and EMS runs, as well as those who received naloxone during an EMS run.
Table 1: Summary of ED visits matched to EMS runs. Note a general increase in naloxone use.

<table>
<thead>
<tr>
<th>Year</th>
<th>Matched ED visits</th>
<th>Unique patients with ED visits</th>
<th>Matched ED visits w/ naloxone</th>
<th>% Matched ED visits w/ naloxone</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>90,003</td>
<td>65,847</td>
<td>1,147</td>
<td>1.27</td>
</tr>
<tr>
<td>2015</td>
<td>86,671</td>
<td>63,491</td>
<td>1,164</td>
<td>1.34</td>
</tr>
<tr>
<td>2016</td>
<td>87,530</td>
<td>64,804</td>
<td>1,705</td>
<td>1.95</td>
</tr>
<tr>
<td>2017</td>
<td>67,462</td>
<td>50,279</td>
<td>1,522</td>
<td>2.26</td>
</tr>
</tbody>
</table>

Discussion

We achieved two important aims in this pilot. First, we enabled a real-world reproducible process for securely hashing, matching, and transferring de-identified data among two or more distinct institutions. Second, we created a unique dataset linking ED encounters from an HIE to their corresponding EMS run from DHS. The benefit of linking the EMS and ED data allows for a more comprehensive understanding of the care continuum, beginning with EMS pickup and continuing through the ED encounter, and will enable future discovery and innovations. (Figure 2)

![EMS Care Continuum](image)

Information such as naloxone administration, which is often not captured in a systematic fashion in the ED, can be used to supplement clinical observations seen in the ED. Alternatively, organizations such as Indiana DHS can improve situational awareness with a better understanding the dispositions for EMS run patients receiving ED care. Enabling the linkage of these two data sources can support a host of research questions aimed at identifying the needs and characteristics of EMS/ED patients across the state of Indiana. Specific emphasis will be placed on examining the readmission rates and clinical outcomes of those individuals who received a naloxone administration during an EMS encounter. We will also refine this generalizable framework to link and share additional data as well as align the technical infrastructures supported by both MPH and the Regenstrief Institute. Our future work will support further integration and analysis of other clinical and non-clinical datasets to support research efforts including the opioid epidemic, precision health, and beyond.

Conclusion

It is feasible to develop collaborations and technical infrastructure to integrate clinical and nonclinical data using reproducible, deidentified, and scalable approaches.

References

Abstract

The Cox proportional hazards model is a popular semi-parametric model for survival analysis. In many situations, data of the same cohort of patients might reside at different institutions. We aim at developing a federated algorithm for the Cox proportional hazards model over vertically partitioned data. Our algorithm, namely VERTICOX, is to obtain the global model parameters in a distributed fashion without collecting individual patient-level data.

Introduction

Survival analysis is widely used for time-to-event healthcare data and can facilitate the comparative study of treatment effectiveness and outcomes of interest (i.e., mortality, disease, recurrence). There have been a lot of studies including individualized survival analysis to evaluate prognostic factors in oncology. Among survival analysis studies, Cox proportional hazards [1] is one of the most popular models. It utilizes the hazard function to assess the significance of covariates to the survival times of individuals or subjects. The model can exploit covariate information to determine the importance of predictors, establish flexible and robust regression models considering time-varying effects, and examine the variation of treatment effects across multiple institutions, among other applications. To make full use of Cox model, we would like to integrate all pieces of fragmented healthcare information associated with patients (i.e., different visits, different sources, different modality, etc.) However, inter-institutional exchange of patient-level data is often restricted due to institutional policies, legislation, or privacy concerns. For example, in the United States, patient data shall be hosted by covered entities in HIPAA-compliant environments [2] without disclosure to unauthorized third parties. The NIH Genomic Data Sharing (GDS) policy [3] requires participants’ consent to exchange genomic data. Effective May 2018, the General Data Protection Regulation (GDPR) [4] of the European Union (EU) requires institutions collecting or processing personal data of EU citizens to protect their privacy. Analysis and prediction methods that do not explicitly exchange patient-level data can enable distributed data analyses in many practical applications. Federated data analysis methods allow researchers to build global statistical models without sharing patient-level data with other parties.

There is a need to develop a survival analysis model using distributed data across multiple institutions. Prediction and analysis of diseases and effectiveness of treatments can be enhanced by making use of massive healthcare data collected by numerous organizations such as hospitals, research institutions and government agencies. For example, the patient-centered Scalable National Network for Effective Research (pSCANNER) [5] integrates data from over 31 million patients to facilitate comparative effectiveness researches and prospective analyses, without sending those data out of their institutions. In this work, we will introduce a recently proposed method, called VERTICOX, for distributed survival analysis over vertically partitioned data, which is based on the alternating direction method of multipliers (ADMM) framework [6]. Distributed optimization is made for the federated data analysis over covariates that are vertically distributed across multiple institutions. A client-server architecture is established for iterative optimization for model parameter estimation.

Method: In the proposed framework, each institution only transmits aggregated intermediary results to the server at each iteration while the individual patient data are kept private within their host institutions (see Figure 1). Distributed model parameter estimation for the Cox proportional hazards model is developed using the Breslow’s partial likelihood function [7] (see Figure 2) based on the aggregated intermediary results. Thus, model parameter estimation and subsequent survival estimation can be conducted in a privacy-protecting fashion. VERTICOX is demonstrated to
converge at a linear rate under the ADMM framework. The proposed method is evaluated on real-world applications over the Internet. Experimental results demonstrate that the proposed method can achieve equivalent accuracy in the estimation of model parameters and statistics when compared to its centralized counterpart.

![Diagram of Breslow's partial likelihood function](image)

**Figure 2**: Estimation of Breslow’s partial likelihood function can be decomposed into a set of linear problems by ADMM

**Results**: The SEER dataset ([https://seer.cancer.gov/data/](https://seer.cancer.gov/data/)) was sampled from 55,007 breast cancer patients with first diagnosis from 2001 to 2009. Twenty features were derived from 9 covariates including age at diagnosis (AAD), race, marital status, histology, grade, tumor size (TS), number of nodes examined (NNE), number of positive nodes (NPN) and estrogen receptor (ER) status using dummy coding. The training set for model parameter estimation involved 1,726 records for 97 distinct event times and the test set consists of 1699 records with 20 features randomly extracted for survival analysis. Table 1 shows the time cost and estimation error for the collaboration between 2 and 3 institutions with respect to its centralized counterpart.

<table>
<thead>
<tr>
<th># of institutions</th>
<th># of iterations</th>
<th>Time cost (seconds)</th>
<th>Estimation error</th>
<th>Server</th>
<th>Train Client</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>1.03 x 10^{-4}</td>
<td>5.37 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>4.65 x 10^{-4}</td>
<td>2.22 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>9.30 x 10^{-4}</td>
<td>2.22 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>3.24 x 10^{-4}</td>
<td>2.22 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>2.60 x 10^{-4}</td>
<td>2.22 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>9.00 x 10^{-4}</td>
<td>6.24 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>6.17 x 10^{-4}</td>
<td>4.64 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>3.56 x 10^{-4}</td>
<td>6.44 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>5.29 x 10^{-4}</td>
<td>8.26 x 10^{-4}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>2.26 x 10^{-3}</td>
<td>1.12 x 10^{-3}</td>
<td>2.10 x 10^{-9}</td>
<td>2.10 x 10^{-9}</td>
<td></td>
</tr>
</tbody>
</table>

**Discussion**: Our results show the proposed method can achieve equivalent accuracy in distributed computation to that of its centralized counterpart. However, the model still has some limitations, especially for very high dimensional data scenarios, in which numerical instability will become challenging and sometime the solver cannot determine unique solutions. Furthermore, as a future research direction we plan to develop a verification mechanism to assess the integrity and faithfulness of the model.

**Reference**

Design and Implementation of an Electronic Survey for Follow-Up of Acute Conditions in Primary Care

Vijan Joshi, MD, John Schrom, MPH, Kyle Munkittrick, MA, Carlo Stearns, MS, Svetlana Ivanova, MSIST, David Hoang, Lenard I Lesser, MD, MSHS, Tamer Fakhouri, MD, Andrew Diamond, MD, PhD
One Medical, San Francisco, CA

Introduction: Acute conditions make up a large portion of the reason for visits to primary care practices1. These acute conditions can sometimes progress to more serious illnesses. Resource limitations make scheduling follow-up visits, calls or messages to check for resolution of these conditions impractical. Awareness of patient outcomes helps prevent clinicians from making unintended diagnostic errors, and engaging patients to seek follow-up care if they are not improving creates a diagnostic safety net2,3. This study aimed to assess provider and patient engagement in an electronic follow up survey after a primary care encounter for an acute condition.

Methods: We identified seventy acute primary care conditions based on their frequency and expected clinical resolution time. When these problems were added to the electronic health record (EHR) of a national primary care system, an electronic survey order was automatically created and sent to the patient at a predetermined follow up interval unless the clinician opted out of the order. The survey (see inset) asked the patient if their condition was now resolved, improving, unchanged or worse. Patients whose conditions were unchanged or worse were prompted to send a message to their provider or schedule a follow-up visit.

We analyzed new, acute problems created during the four week period following the initial roll-out of the surveys. We explored provider opt-out of and patient responses to the surveys, along with attributes of the provider, patient, and condition. Using stepwise logistic regression, we examined the association of provider characteristics and condition type with the likelihood of provider and patient engagement with the feature. Analyses were performed in R v3.4.1 with MASS v7.3.

Results: The sample consisted of 20,402 acute problems created in the four weeks after January 25, 2019. Providers opted out for 43% of the problems, allowing 12,809 follow up surveys to be sent. Once sent, the patient response rate was 46%. Most often, patients indicated the problem was either resolved (28%) or improving (54%), while fewer patients responded saying it was unchanged (16%) or worse (2%). Select results from preliminary models are shown in the table. Provider characteristics seemed to play a substantial role in opt-out behavior, with providers who had been at the practice longer and those who had longer panel waitlists being less likely to opt-out of follow-up. Provider characteristics, including having a primary care relationship with the patient, did not seem to influence patient engagement. The specific condition seemed to impact both provider and patient behavior. Providers more consistently opted out of sending surveys for musculoskeletal-related pain, gastrointestinal issues and vaginal concerns, but were less likely to opt out of sending surveys for acute respiratory problems. Patients seemed to respond more consistently to requests for follow-up on their gastrointestinal issues and some respiratory complaints, but were less likely to respond to requests related to certain musculoskeletal pain.
**Discussion:** Electronic surveys appear to be a useful method for collecting information about patient progress and the natural course of acute conditions. Initial models begin to paint a picture of testable hypotheses and potential caveats that can be further investigated as use of this feature grows.

The provider engagement characteristics evaluated here suggest opportunities for improving change management and mentoring of newer providers in the use of this feature. It also suggests that the type of condition may influence engagement in the feature, as providers may not seek follow-up for certain conditions, and patients may similarly not respond. These biases may be concordant, as seen with musculoskeletal pain, or discordant, as with gastrointestinal complaints. There are multiple hypotheses for why this may be: differences in treatment plans based on the type of condition, provider or patient comfort regarding the topic, or incorrectly calibrated follow-up criteria or timeline. Further study is warranted to understand and mitigate these phenomena.

The utility of this new data source is exciting: having contemporaneous progress data about acute conditions opens doors for direct and timely follow-up, clinical phenotyping of acute conditions and smarter clinical decision support. These initial results find that patients and providers largely embrace this tool for many situations. Further study is underway to understand patterns of provider adoption and condition-based variations in provider and patient engagement.

<table>
<thead>
<tr>
<th>Provider Opt-Out, AOR (95% CI)</th>
<th>Patient Engagement, AOR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider saw patient in panel</td>
<td>1.185 (1.273 - 1.367)</td>
</tr>
<tr>
<td>Provider tenure at practice (years)</td>
<td>0.948 (0.934 - 0.962)</td>
</tr>
<tr>
<td>Provider panel waitlist (n)</td>
<td>0.999 (0.998 - 1.000)</td>
</tr>
<tr>
<td>Problem: Acute Bronchitis</td>
<td>0.621 (0.484 - 0.795)</td>
</tr>
<tr>
<td>Problem: Diarrhea</td>
<td>1.323 (1.057 - 1.656)</td>
</tr>
<tr>
<td>Problem: GI Disturbance</td>
<td>1.666 (1.022 - 2.717)</td>
</tr>
<tr>
<td>Problem: Shoulder Pain</td>
<td>1.206 (0.985 - 1.476)</td>
</tr>
<tr>
<td>Problem: Lower Back Pain</td>
<td>1.388 (1.188 - 1.621)</td>
</tr>
<tr>
<td>Problem: Knee Pain</td>
<td>1.401 (1.172 - 1.675)</td>
</tr>
<tr>
<td>Problem: Pneumonia</td>
<td>--</td>
</tr>
<tr>
<td>Problem: Vaginal Discharge</td>
<td>1.929 (1.284 - 2.896)</td>
</tr>
</tbody>
</table>

**References**

Creating a Database for Health IT Event Reports by Using a Hybrid Deep Learning Model

Hong Kang, PhD, Yang Gong, MD, PhD
School of Biomedical Informatics, the University of Texas Health Science Center at Houston, Houston, Texas, USA

Introduction

Health information technology (Health IT, HIT) has been widely applied in the clinical settings in the forms of electronic health record (EHR) systems, administrative or practice management systems, automated dispensing systems, laboratory information systems, diagnostic imaging systems, etc. Safe use of HIT holds promise in helping healthcare system save costs, improve patient outcomes, decrease occurrence of medication errors, and refine healthcare process measures across diverse settings. However, unavailability, malfunction, and improper use of HIT have created new risks in practices which encourage workarounds and increase the likelihood of adverse events.

Listed in the top 10 technology-hazards for healthcare and comprised of one sixth all patient safety events, HIT events are a major concern for patient safety. Collecting reports of HIT events is essential for aiding the understanding, characterization, discovery, and reporting of HIT events toward improved patient safety. The Agency for Healthcare Research and Quality (AHRQ) has spearheaded this task by developing a set of standardized event reporting Common Formats (CF). In the latest CF Hospital Version 2.0, HIT is considered a set of contributing factors rather than as a specific event category. This modification highlighted the significance of HIT factors in patient safety events yet increased the difficulty of archiving HIT event reports in Patient Safety Organization (PSO) institutes. Usually, reporters prefer to leave HIT fields blank due to lack of knowledge in patient safety event reporting, which greatly hinders the accumulation of valuable reports and blocks the opportunity to learn from previous events. Therefore, a rich resource for HIT events is urgently needed for the patient safety community.

We previously proposed a review strategy to identify HIT events from the US Food and Drug Administration (FDA) database, Manufacturer and User Facility Device Experience (MAUDE), based on a keyword filter and sociotechnical model. The review identified 3,521 HIT event reports and revealed that MAUDE holds up to 50,000 HIT event reports, and is currently the most abundant resource of HIT events. To assist manual review, we further applied TF-IDF and topic modeling to develop a classifier, which exhibited a relatively low recall of the identified reports. Recently, the application of advanced artificial intelligence algorithms has shown promise in advancing biomedical research. In this study, we assessed 9 individual and 120 hybrid models that are commonly used in the deep learning community in the identification task of HIT events. By applying the optimal hybrid model on all MAUDE reports (1993-2018), 48,997 reports were identified for establishing the first database of HIT event reports.

Methods

Data source. 6,994 reviewed reports (narrative), including 3,521 HIT event reports and 3,473 non-HIT event reports were identified from MAUDE database from 2008-2016. The dataset was split into a training set (70%), development set (10%), and test set (20%). Individual models. Nine individual models were developed for the identification of HIT event reports, including four classic machine learning models (SVM, logistic regression (LR), naive Bayes (NB), and random forest (RF)), and five popular deep learning models for document classification (CNN, RNN with/without attention layer (RNN, RNN_att), and Hierarchical RNN with/without attention layer (HRNN, HRNN_att)). TF-IDFs were calculated and applied as the report features of SVM, LR, NB, and RF models, while word vectors derived from Global Vector for Word Representation (GloVe) were applied as the features of deep learning models. Hybrid models. The best seven individual models from the nine were selected for hybrid model construction. 120 hybrid models (all possible combinations of the seven, i.e., \( C_7^2 + C_7^3 + \cdots + C_7^7 \)) were constructed by concatenating the output layers and all models were evaluated to determine an optimal hybrid model. Second-round evaluation. The MAUDE 2017 database (935,715 raw reports), independent from the training, development, and test sets, was used for a second round evaluation. The raw data was pre-processed by a keyword filter before being fed to the optimal hybrid model. Six domain experts in two groups reviewed 850 (~10%) reports sampled from the 8,462 filtered reports to identify HIT events. Google universal-sentence-encoder was applied in the sampling task to ensure the semantic diversity of sample reports. Divergences among the reviewers were fixed through group discussion. The review results were used to further assess the optimal hybrid model. Database for HIT events. We applied the keyword filter and the optimal hybrid model on all MAUDE reports from 1993 to 2018 (7,903,397 reports) to establish a database for HIT events.
Results

We used accuracy and $F_1$ score on the test set to measure the model performance. NB and RF models were excluded from the construction of hybrid models due to their weak performance. Figure 1a indicates that 1) deep learning models generally outperform classic models, and CNN ($\text{Acc}=0.892, \text{F}_1=0.864$) has the best performance among individual models; 2) most hybrid models outperform individual models; 3) the hybrid model (the purple dot at top right corner) which consists of CNN, HRNN, and LR ($\text{Acc}=0.902, \text{F}_1=0.876$) is optimal. The second-round evaluation shows that the optimal model has a ~0.862 accuracy on MAUDE 2017, an independent dataset (Figure 1b). The accuracy stabilizes after evaluating 400 most diverse reports, so the sample size of 850 is sufficient for this evaluation. With the help of our model, we established the first HIT event database, containing 48,997 reports identified from the MAUDE database of all years, from 1993 to 2018 (Figure 1c).

![Figure 1.](image)

Discussion

Learning from HIT events is challenging due to the lack of specific databases. This study provides a solid strategy to identify HIT event reports from a public database, and has established the first database for HIT events. The strategy is intended to organize the events by connecting and synchronizing them with large databases, which would help reporters and reviewers describe, understand, and integrate the events connecting to a fuller spectrum of HIT events. The database will help researchers conduct root cause analysis and develop learning tools or systems to improve the safe use of HIT. Additionally, manufacturers can use the database to improve the surveillance of product quality. We performed a systematic comparison of classic machine learning models and popular deep learning models. All possible combinations of individual models were evaluated, and an optimal hybrid model consisting of three individual models (CNN, HRNN, and LR) was proposed. With the superiority of deep learning, the hybrid model has the potential to be applied on other datasets with narrative reports. An extended training set involving the resources from FDA, PSO, and AHRQ PSNet could improve the applicability of our model and facilitate the data interchange among healthcare institutes. The hybrid model also has the potential to identify reports of other event subtypes such as patient falls and medication errors.

Conclusion

We proposed a hybrid deep learning model to provide a standardized process for identifying HIT events from an FDA database. The hybrid model helped identify 48,997 HIT event reports, which verified our prior estimation in establishing the first database for HIT events.

References


Nitu Kashyap, MD¹, Allen Hsiao, MD¹

¹Yale New Haven Health and Yale School of Medicine, New Haven, CT

Abstract

Physician burnout and distress is a growing problem due to complex combination of loss of autonomy, increased administrative tasks, and inefficient work environments. The EHR and information technology (IT) systems are often cited as major contributors. Decreasing cognitive load across multiple areas of the EHR can have a cumulative impact in improving provider efficiency and experience with the EHR.

Introduction

Attention and the realization of the importance of “physician burnout” has been growing the past several years. Studies are recognizing that physicians, and clinicians in general, are facing increasing pressures and stress, increasingly “ugly” stress, due to imbalance of resources, lack of control, and lack of meaning.[1-3] This is due in no small part to a “perfect storm” of increasing financial pressures, shorter patient visits, and compliance, regulatory, & payer documentation requirements, the EHR is the hammer often used to meet those burdens. The loss of autonomy and control and increased pressures for higher productivity, at least by patient volume, are significant contributors to feelings of burnout. While there are many different facets and components to physician burnout, and therefore requisite different strategies, we will focus on some of the key information technology approaches for mitigating burnout. Locally, we have dubbed this strategy, “Less time typing, more time caring.”

Methods

We implemented a toolkit of Information Technology interventions and evaluated the impact of each including:

1. Secure Messaging Application: Introducing a secure, HIPAA compliant messaging application that also serves as provider directory, makes it easier for users to communicate efficiently, in a secure manner, while decreasing the time previously spent in playing ‘phone tag’.


3. Speech Recognition: Physician satisfaction is very high for speech recognition. [5, 6]. A phased rollout of speech recognition software, integrated with the EHR allows providers to use a combination of voice-to-text, keyboard shortcuts, and EHR shortcuts to improve documentation time and quality.

4. EHR Integration with State prescription monitoring program (PMP): CT State mandates checking PMP information for every patient who is prescribed an Opioid. This required logging into a separate system, searching for patient and scrolling for information. Creating a single sign on integration with patient context, we improved access to this vital information, eliminating barriers and saving time.

5. Clinical Decision Support (CDS) Curation: By creating dashboards and tools to monitor utilization of CDS alerts, and establishing an oversight committee we were able to retire alerts, modify others, and change decision support to other non-interruptive modality in several areas.

6. Virtual Scribes: Juggling the keyboard, mouse, microphone, navigating EHR is a cognitive burden that hinders patient-physician interaction. Scribes can document in real-time, and navigate the chart via voice commands. Working with technology vendor, transcription experts, EHR vendor, and compliance, we created a program where a scribe is available to providers via secure phone connection, and can document in the EHR concurrently during the visit.
Results

To date, results show promising progress made by each IT strategy to save time or improve clinical workflow and decrease time spent at the computer.

Figure 1: Impact of targeted IT strategies on reducing physician burnout

Discussion

EHR and supporting systems are understandably often viewed as barriers to patient care due to design and implementation decisions made to support billing, compliance, and data security. Informatics plays an important role in mitigating these frustrations and contribution to physician burnout. Individually, each tool may not be as impactful, but as a collective strategy it has been noticed and appreciated. Highest satisfaction and perceived impact has been from speech recognition and virtual scribe program. Making it easier to do the right thing, there is decrease in unsafe workarounds like copy/pasting in notes, using phones to text colleagues about patient care inadvertently including PHI, canceling important alerts and warnings by habitual action. A common theme across all these applications is to decrease cognitive load, remove meaningless tasks to background, while allowing providers more time to focus on patient and patient care needs. Less time typing, more time caring.

Future Direction

We continue to evaluate Technology and Informatics methods to mitigate inefficient workflows, clicks, and time spent in the EHR including expansion of virtual scribe program, Piloting cloud speech recognition solution, Review & improve security log off timings, Continue CDS curation to decrease alert fatigue, and Advocate for billing reform.

Conclusion

While EHR and electronic systems are perceived as a major contributor to burnout, information technology solutions can also help mitigate or lessen the impact. Continued evolution and application of mitigating solutions is critical in supporting physicians facing more pressures than ever before.

References

Ontologies Enabling Computable Tables

Jeffrey G. Klann, PhD\textsuperscript{1,2,3}; Nich Wattanasin, MS\textsuperscript{2}; Michael Mendis\textsuperscript{2}; Matthew Joss, MS\textsuperscript{2}; Hossein Estiri, PhD\textsuperscript{1}; Kavishwar Waghholikar, MBBS, PhD\textsuperscript{1,3}; Shawn Murphy, MD, PhD\textsuperscript{1,2,3}

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

Introduction

Machine learning algorithms are becoming a standard tool for retrospective analysis of clinical data. Such algorithms frequently expect input data in a sparse multidimensional array, aka a very wide table. The columns in the input dataset are called “features” and represent curated knowledge, such as “date of first diagnosis of diabetes mellitus”. We will call such sparse arrays “computable tables”.

A structural transform is needed, as most clinical data warehousing models use narrower tables with one row per record (e.g., one row per diagnosis, lab result, or observation). But, developing such a computable table is also predicated on a deep understanding of coding systems and value sets. Typical clinical data is made up of standard codes (e.g., ICD), proprietary codes (e.g., Epic procedure codes), and local codes (from customized local systems). Codes are retired and added on a regular basis. Additionally, some of the knowledge is local to an institution - both the local codes and the knowledge of what codes are used (e.g., RxNorm ingredient vs dose-forms). These challenges necessitate some type of knowledge-management system.

Moreover, it is frequently possible to describe these codes in terms of hierarchical groups, and efficient grouping at varying levels of granularity is essential for computable tables. For example, to develop predictive algorithms of disease using co-morbidities one might want a high-level feature like “cerebrovascular disease”, whereas when predicting outcomes after cerebral infarction, one might want much more granular diagnostic features, such as “cerebral infarction due to thrombosis of cerebellar artery.”

A variety of projects (such as UMLS, BioPortal, OHDSI, PhcKB, VSAC) handle the knowledge management and/or grouping of standardized code content, to varying degrees. However, the next step of generating feature-oriented computable tables is performed by hand, with manual per-project programming to combine curated code dictionaries with valuesets and analytic data models.

Here we propose two approaches to automate production of computable tables for machine learning algorithms by using curated hierarchical ontologies. We believe these are straightforward enough for use by researchers and data analysts. We discuss our preliminary results implementing and deploying these approaches at Partners Healthcare.

\textit{i2b2 Ontologies}

i2b2, now in its fifteenth year and in use at over 200 institutions worldwide, is the eldest tool for data exploration on large clinical data warehouses. \cite{1} It was designed with a robust ontology system as an organizing principle. Folders of codes are arranged hierarchically, and many standard hierarchies (via BioPortal, OHDSI, and UMLS for example) have been imported to i2b2 and can be augmented with local codes added into the hierarchies.

i2b2’s ontology is primarily seen in the web-based cohort-finding tool, but the robust API makes it programmatic access possible as well. Our approach leverages the i2b2 ontology system, which is implemented at Partners Healthcare to organize local and standard terminologies in a 3-million patient enterprise data warehouse.

\textbf{Method}

At times, researchers are studying problems that are sufficiently well-defined so that the desired computable table can be described at the time of the data request. At other times, researchers desire the flexibility of having all of the detailed data, but in this case, they need tools to arrange new computable tables as the problem becomes better understood. To support both cases, we have defined a pre-computed and just-in-time approach.

\textit{Pre-computed}

We developed an approach and user experience that allows users to interactively define a computable table by specifying groupings driven by ontology definition (e.g., as general as “Gastrointestinal disorders” or as specific as “Crohn’s Disease of the small intestine”). We identified the following groupings, based on our experience with data requests at Partners Healthcare: by date (date of first occurrence, or date of most recent occurrence), by existence (presence/absence, or count of occurrences), and by code (most frequent code, or list of all codes).

We implemented this approach as an i2b2 webclient plugin. It allows users to, starting from a previous cohort query, add ontology items, select the information groupings, and export the data in CSV format for use in analytical tools.

\textit{Just-in-time}
Inevitably, researchers ask for more detailed data than is available in the pre-computed table, because it is often preferable to slice-and-dice the data in different ways as the research methods mature. For this use-case, our approach is to provide users access to subsets of data in the source data warehousing format and augment this with tools for analysts to later perform the same type of ontological groupings described in the pre-computed approach. At Partners Healthcare, this source format is the raw i2b2 “fact” table. This is a long, narrow table in which each row represents a single, specific fact about the patient. Such an extract has limited utility on its own, but when combined with an ontology, hierarchical organization and grouping become possible.

We implemented this approach with an i2b2 bulk-export plugin and an additional plugin that provides ontology metadata and SQL code for generating code sets for groupings. For example, if a researcher drags ‘Diabetes Mellitus’ into this plugin, she will retrieve a SQL code snippet to build a value set of all codes (both standard and local) in the local database that represent that disease. (See Figure 1.)

In combination with the fact table export, a data analyst can use these code sets to create their own analytical views of these data that can be output to machine learning algorithms.

Results
Partners Healthcare plans to deploy these plugins on its enterprise data warehouse, a 3-million-patient dataset already organized around a customized i2b2 ontology. This “Partners Data Enclave” will be a sandboxed analytic environment for investigators.

An Enclave pilot was recently completed as part of a machine learning challenge. Teams created algorithms to solve problems using machine learning and statistical algorithms. Teams registered to access clinical data for the over 80,000 Partners Biobank patients. The teams were given access to the Biobank Portal, an enhanced i2b2 query tool for cohort design with genomic and clinical data. [2] Our plugins were debuted in this environment (see Figure 1).

At present, development of the Data Enclave environment for the full enterprise data warehouse is underway. The tools and dataset are being adapted for this significantly larger endeavor. A pilot test of the Enclave, studying inherited heart disease, is being conducted concurrently with development.

Discussion
Moving data into data warehouses has been the focus of much work (Extract, Transform, and Load tools), and cohort-finding analytic tools are becoming mature. However, a critical step in the analytic pipeline is to structure and organize the data for analysis by external tools. The optimal characteristics and methods to create computable features in these wide, sparse, curated tables are not well well-studied. For example, a feature must be specific enough to have some predictive power yet must be frequent enough to have statistical significance.

We propose approaches that leverage the mature, well-tested capabilities of hierarchical biomedical ontologies to enable researchers to easily build these tables. We demonstrate this in the mature i2b2 platform, where we provide two approaches to support researchers’ development of computable tables.

Conclusion
Data warehouses must be combined with hierarchical biomedical ontologies to develop the curated, complex computable tables needed by machine learning algorithms. Here we demonstrate tools that support researchers in this task using i2b2. We are pilot testing these tools in a machine learning challenge at Partners Healthcare.

Acknowledgements
Thanks to the i2b2 and Biobank Portal teams. Funded in part by NIH NHGRI RO1 HG009174 (“Developing i2b2 into a Health Innovation Platform for Clinical Decision Support in the Genomics Era”)

References
Identifying Early Hepatic Encephalopathy Through Digital Phenotyping
Jejo D. Koola, MD, MS,1 Veeral Ajmera, MD,1 Job Godino, PhD,2 Lauren L’Heureux,1 Amanda Gooding, PhD,3 Marc Norman, PhD,3 Faraz Hussain, MS,4 Alex Leow, MD, PhD4
1. Department of Medicine, UCSD, San Diego, CA
2. Department of Family Medicine and Public Health, UCSD, San Diego, CA
3. Department of Psychiatry, UCSD, San Diego, CA
4. Department of Psychiatry, University of Illinois at Chicago, Chicago, IL

Background
Cirrhosis, irreversible scaring of the liver, causes decreased mental, physical, and biochemical function. The prevalence is estimated between 400,000 and 3,000,000 persons in the United States, and the disease causes 44,000 deaths annually.1 Hepatic Encephalopathy (HE), a potentially reversible neuro-psychiatric complication of cirrhosis, causes significant morbidity and mortality. Symptoms range from inattentiveness and mild personality changes to coma. Hepatic Encephalopathy frequently leads to hospital admission if uncaught and left untreated, costing $1.6 billion to $2.0 billion annually.2

Symptoms may rapidly fluctuate with a time course of hours to days. HE is frequently divided into various stages depending on severity. HE is classified into two broad categories: “Covert” and “Overt” HE. Symptoms of covert HE include inattentiveness, sleep cycle disturbance, and anxiety. Overt HE, which can result in coma, occurs in 30 to 45 percent of patients with cirrhosis, and covert HE is estimated to affect up to 60 percent of patients.3 Patients frequently do not present to the healthcare setting until their disease progresses to overt HE. Though the symptoms of covert HE are protean, they have been associated with increased motor vehicular accidents and is a risk factor for progressing to overt HE.4 Early identification may motivate early treatment; however, diagnosis of covert HE requires complex and time-consuming neuropsychological testing, which is often infeasible in the clinical setting.

A possible solution is the use of passive sensors: activity and exercise monitors, location trackers, heart rate sensors, etc.. The use of passive sensor technology has been tested on detecting mood changes in psychiatric disorders.5–7 This interaction between health and technology has been termed the “digital phenotype.”8 We designed and conducted a feasibility study using mobile health technology and wearable sensors to monitor patient’s activity levels, vital signs, technology interaction, and sleep hygiene as a means of “phenotyping” hepatic encephalopathy.

Methods
We used a prospective cohort design with repeated measures to structure this feasibility study. We recruited adult patients with cirrhosis who were able to use a smartphone when cognitively intact. Exclusion criteria included severe disease with Model for End Stage Liver Disease (MELD) score > 30, uncontrolled neuropsychiatric diagnoses, psychoactive medications apart from stable antidepressants, red-green color blindness, and active drug or alcohol abuse.

The primary outcome measure was cognitive function as assessed by a battery of neuropsychological testing. Cognitive testing occurred monthly in a controlled environment by a trained professional and included three tests: the Psychometric HE Score,9 the Inhibitory Control Test (ICT),10 and the Encephalapp Stroop test.11 The Encephalapp test, a standard Stroop test, measures the time for a subject to complete an activity in the presence of a distractor. Normal ranges for subjects with intact cognition range from 120 seconds to 150 seconds depending on age, gender, and education. Additionally, we measured cognitive status in the home setting weekly by having subjects perform the Encephalapp Stroop test on their smartphone.

We combined multiple digital datastreams from patients with advanced liver disease collected from the patients’ smartphone utilizing the BiAffect app,7 which unobtrusively and continuously collects typing speed, typing accuracy, and accelerometer data. Additionally, we fitted subjects with a wearable activity monitor (FitBit Charge 2, FitBit, Inc.) for sleep, activity level, and heart rate measurement. We collected basic demographic data including age, biologic sex, and education level and clinical data relevant to patients with advanced liver disease.

Figure 1: Overall study design for digital phenotyping of hepatic encephalopathy.
As this was primarily a feasibility study, we did not attempt to power this study for hypothesis testing. Instead, we focused on identifying barriers to recruitment, subject acceptance, and technical integration. In this report we qualitatively report our results.

**Results**

We evaluated 24 subjects for participation, 15 were excluded based on chart review. Reasons for exclusion on chart-review included: high MELD score (4), active drug or alcohol abuse (9), and severe extra-hepatic disease (2). Nine subjects were approached and 4 were excluded because they did not own an iPhone and 3 declined to participate. Reasons to not participate included: insufficient compensation (1) and no information given (2).

Presently, two subjects are actively enrolled. Average age was 42 ± 4 years. Average MELD score was 19 +/- 3. Causes were viral hepatitis in one case and Budd-Chiari syndrome in the second. Eighteen total neuropsychological assessments were performed between the study clinic and the home setting. Thirteen of eighteen assessments indicated that the subject had covert hepatic encephalopathy. Of a total of 211 potential patient-days (patients x days) of data collection, subjects’ smartphones and FitBits submitted 114 patient-days of data. Subjects did not report any adverse events related to the study. The greatest amount of missingness occurred in collecting sleep data from the associated FitBit. Figure 2 depicts selected correlations of the Encephalapp Stroop score against several of the predictor variables.

**Discussion**

Improving care for hepatic encephalopathy will require novel care models including approaches that leverage technology. Early hepatic encephalopathy often remains undiagnosed partly due to underreporting as well as under-ascertainment. In this study we demonstrate the feasibility of a digital phenotyping platform that collects continuous data from multiple datastreams including keyboard kinematics from a smartphone and activity and sleep patterns from a wearable activity monitor. In this feasibility trial, barriers to enrollment included rigorous inclusion/exclusion criteria and lack of an iPhone (the BiAffect app currently only runs on the iOS platform). Technical challenges included data integration from datastreams with widely different scales and measurement frequencies as well as integrating disparate technological platforms. Though this study is not powered to formally test association, trends between the digital phenotype markers and measures of cognitive function exist (Figure 2). Future directions include expanding recruitment protocols and implementing robust methods of analyzing temporally varying data.

**References**

Expanding the Precision Medicine Landscape: How to Prepare Providers Who Serve Underserved Minority Patients

Rita Kukafka, DrPH, MA1, Margaret Sin, MA1 Thomas B. Silverman, MPH1, Katherine D. Crew, MD, MS2

1Department of Biomedical Informatics, Columbia University, New York, NY; 2Department of Medicine, Columbia University, New York, NY

Introduction

A goal of precision medicine (PM) is to replace the “one-size-fits-all” approach with customized preventive and therapeutic options that take individual genetic, environmental and lifestyle variability into account.1, 2 While the implementation of PM into clinical practice has been slow3, 4, the Cancer Moonshot Blue Ribbon Panel recommends a focus on what can be done now, even without new scientific breakthroughs.5 Embedded in this recommendation, is a call for increased testing for hereditary cancer syndromes in individuals with an increased likelihood of having a pathogenic variant. Such testing allows those identified as mutation carriers to begin early cancer prevention.

Hereditary Breast and Ovarian Cancer (HBOC) syndrome is among the most common of the inherited cancer syndromes. Women with HBOC have a substantially higher breast and ovarian cancer risk than those without susceptibility mutations, with a cumulative risk of developing breast and ovarian cancer of up to 80% and 40%, respectively. Identifying women with genetic risk provides an opportunity to offer personalized cancer risk management strategies. Currently available strategies can decrease breast and ovarian cancer incidence by up to 95% and 90%, respectively.6, 7 However, for such benefit to be realized, at-risk individuals need to be identified. Fewer than 15% of women at risk for HBOC are referred for genetic testing; less among women from racial/ethnic minorities. Most individuals who receive genetic testing for BRCA 1/2 mutations are white, well-educated, and have high incomes. Providers that serve minority patients are less likely to identify and refer patients for genetic testing.8

Major obstacles to increasing genetic testing for HBOC and downstream care in minority populations include patient lack of awareness, financial, and competing life concerns. Provider-level barriers include low levels of perceived clinical value of testing; gaps in genomic knowledge and skills; and other competing clinical and educational priorities.9

Our team developed the RealRisks Patient Decision Aid and the BNAV Provider Decision Support system to simultaneously address both patient and provider-level barriers to expanding genetic testing for HBOC. These tools are integrated into the clinical workflow of outpatient clinics of Columbia University/New York Presbyterian Hospital in New York, which serves a large population of low-income and minority patients. These tools are currently being evaluated in a cluster randomized trial.10 Eighty providers who have enrolled in the trial completed a baseline survey to assess preparedness for implementing genetic testing and managing high-risk breast cancer patients. While previous surveys have been conducted, little is known about providers that serve minority low-income patients.

Methods

Among primary care providers (N=80), including internists, family practitioners, and obstetrician-gynecologists (OB/GYN), we administered a survey which assessed demographics, confidence in managing patients with family history of breast or ovarian cancer, knowledge of predictive genetic testing for breast cancer, orientation towards shared decision making, attitudes towards making a referral for genetic testing (ATT), subjective norms (SN)—i.e., how providers feel that patients or other providers expect them to make decisions about genetic testing referral—perceived behavioral control (PBC) in making a referral to genetic testing, and lastly, behavioral intention (BI) of making a genetic testing referral. We measured ATT, SN, PBC, and BI variables using 7-point Likert scales. Confidence was measured using a scale of 0 to 4 where a higher score indicated higher confidence. The last four constructs (ATT, SN, PBC, and BI) are drawn from the Theory of Planned Behavior, which posits that ATT, SN, and PBC can be used to explain an individual’s BI, which in turn can predict an individual’s behavior. We examined whether there were any differences observed between obstetrician-gynecologists (OB/GYN) and non-OB/GYN providers. In order to detect possible differences between provider specialties, we ran two-sided Mann Whitney U tests comparing OB/GYNs and non-OB/GYNs. We also generated a descriptive analysis on participants’ baseline characteristics, and fitted the Theory of Planned Behavior constructs into a multiple linear regression model to assess whether ATT, SN, and/or PBC predicted providers’ BI of recommending genetic testing.11
Results

About half of the providers were OB/GYN’s (37; 46.25%) and the rest were from family practice (13; 16.25%) or internal medicine (30; 37.50%). The majority of respondents were female (71; 88.75%), and not Hispanic or Latino (67; 84.81%). Most of the providers had good knowledge of predictive genetic testing for breast cancer and favorable attitudes towards making a referral. Providers were only moderately confident in managing patients with a family history of breast/ovarian cancer (M=2; SD=1.4), and about one third (37%) reported having ever using a breast cancer risk assessment tool. Compared to non-OB/GYNs, the OB/GYNs were more likely to report that they have had patients who requested genetic testing in the past 2 years (p<.001). OB/GYN’s were also more likely to report that patients and other providers expect them to make a genetic testing referral after risk assessment (SN; p=0.035) and that it was easier for them to make a decision about a genetic testing referral (PBC; p=0.005). Among all respondents, we found that the TPB constructs of ATT (p=0.051), SN (p=0.003), and PBC (p=0.099) are significant, positive predictors of BI to make a genetic testing referral.

Discussion

The expansion of PM into primary care settings that serve minority populations will depend on the extent to which providers are prepared in incorporating advances in genomic medicine into practice. Our results demonstrate that while providers held positive attitudes and adequate knowledge about genetic testing, they were only moderately confident in managing patients with a family history of breast/ovarian cancer. The TPB provided useful insight into how providers’ intentions are shaped. Overall, our findings strongly suggest that perceived expectations of patients and peers (SN), together with perceived confidence (PBC), are strongly predictive of intention to make a genetic testing referral. Therefore, efforts to expand genetic testing should help providers become more comfortable in managing high-risk patients and referring them for genetic testing services. They should also support the incorporation of risk-stratified prevention into routine care. Education coupled with skill building resources that are designed to foster confidence in identifying and managing patients with a history of breast/ovarian cancer is needed. Confidence is one of the most important personal factors influencing clinical decision making, and it is also critical in patient-provider communication, which strongly influences utilization, access, and uptake of genetic testing.12 These findings confirm the need for decision support tools like those we are currently evaluating in a clinical trial.

References

Hidden Gaps in Using Common Data Models to Achieve Interoperability Between Electronic Phenotypes and Clinical Data

Fabrício S. P. Kury, MD¹, Li-heng Fu, MD¹, Chi Yuan, MS², Ida Sim, MD, PhD³, Simona Carini, MA³, Chunhua Weng, PhD¹

¹Columbia University in the City of New York, New York, NY, United States; ²Nanjing University of Science and Technology, Nanjing, Jiangsu Province, P.R. China; ³University of California, San Francisco, San Francisco, CA, United States

Abstract

Common Data Models (CDMs) help address the portability barrier of electronic phenotyping algorithms. In this presentation, we share empirical knowledge of how interoperability takes more than what is currently defined in current state-of-the-art data models and how they can be enhanced, leveraged by our experience with manually annotating the free text of the eligibility criteria of 1,000 clinical trials. We discuss how those criteria, as a depiction of clinical researchers’ “real-world” phenotyping needs, highlight three types of difficulties in specifying a phenotype: domain uncertainty, granularity uncertainty, data table uncertainty; and we postulate that these challenges might be addressable by implementing heuristic conventions as a component of the data model.

Introduction

Electronic phenotyping is one of the most pursued secondary uses of the electronic health records data. Prior work has studied phenotype development (how to specify or uncover its normalized codes and workings) and implementation (how to address limitations of observational data), while assuming that selecting what types of patients are worthy of phenotyping is not a concern. Our group has pursued the latter, that is: given a large dataset of types of patients of demonstrable scientific interest, how could they be phenotyped at scale? The referred dataset are the eligibility criteria of recent clinical trials in ClinicalTrials.gov, which we envision could be phenotyped at scale via machine learning capable of translating such free-text into executable database queries. We further assume those queries would be made against a database conformant to the OHDSI OMOP Common Data Model version 6 (hereafter called “the CDM”), as an example of a popular data model with an extensive set of tools. With these premises, our group has annotated the eligibility criteria from 1,000 trials. In this presentation we focus not on this dataset itself, but rather on our findings in performing such task, in order to demonstrate and discuss how clinical researchers’ “real-world” needs cast light onto open challenges yet to be addressed when using a common data model to specify a phenotype.

Methods

Dataset: We extracted from ClinicalTrials.gov the eligibility criteria of a random sample of 1,000 clinical trials of Phase 4 that were actively recruiting as of August 2018, which we then loaded into the brat annotation tool (http://brat.nlplab.org). Two annotators with medical background (FK and LHF) annotated the criteria over 8 months, producing 85 thousand individual annotations.

Annotation: The development of the annotation model followed an informal, interactive process between the two annotators (FK and LHF), both with medical training, and the machine learning developer (CY), upon the shared goal of producing the best possible training data for the machine to learn. It was evident that, whichever intermediate model we used to represent the free text, at some point it would be necessary to translate it into the tables, columns and values needed to compose the intended query. This led to the approach of directly reusing, as an annotation model, the precise definitions of the data model. For example, whatever element in a given eligibility criterion the CDM specifies as belonging to the Measurement Domain, the annotators should demarcate as a Measurement entity; and so forth for all Domains and Data Tables. Upon this basis, other annotation constructs were added to allow for specification of logical operations and relationships between the CDM-based entities.

The CDM: The OHDSI OMOP CDM is designed to hold all observational health data elements that are relevant for analysis use cases in a person-centric schema. At its core are the Standardized Clinical Data Tables, which obey a strict relationship to each of the CDM Domains. As per its documentation, “Domains are identified and separately defined (...) if they have an analysis use case and (...) specific attributes that are not otherwise applicable.” For example, and again paraphrasing the official documentation, “the ‘Condition’ Domain contains Concepts that
describe a condition of a patient, and these Concepts can only be stored in the condition_concept_id field of the CONDITION_OCCURRENCE and [its derivation, the] CONDITION_ENA tables.”

**Results and Discussion**

When attempting to define a phenotype’s database queries from its free text description, some barriers are caused by the nature of the free text, and some by particularities of the desired data model. The former have been explored in the field of Natural Language Interfaces to Databases⁵, could be mitigated by manually rewriting the text with less ambiguous statements, and are not discussed here. The latter are data model-specific, could hardly be addressed by rephrasing unless extensive rewording is accepted, and we discuss below. It is important to clarify that, even for a Phase 4 trial, the needed data might not be present in a “regular” EHR from routine care; however, we considered that what constitutes “routine care” is difficult to define and can vary, and chose to take the following perspective: assuming that the patient information was stored in the EHR in some form, what challenges would one face to query that data? This way, we used the real-world need as a guide for how to approach Medical Knowledge Representation.

**Domain uncertainty** designates when it is unclear what is/are the appropriate domain(s) to assign a given piece of free text, particularly amidst iterative micro-level semantic role labelling⁶. If a criterion reads *Known to be severely alpha-1-antitrypsin deficient*, does *alpha-1-antitrypsin deficient* belong to the Condition domain (SNOMED 30188007) and therefore to the CONDITION_OCCURRENCE CDM table, or is *alpha-1-antitrypsin a Measurement* (SNOMED 88884005) and *deficient* its result (SNOMED 260372006)? In other cases there are also valid codes within the Procedure domain, which adds further uncertainty, for example *HBsAg-positive* (SNOMED procedure code 314099008 *Hepatitis B antigen screening*).

**Granularity uncertainty** describes the unknown preferences between pre- and post-coordination, which is particularly prevalent in a pan-vocabulary data model such as the OHDSI OMOP CDM. For example, if a criterion mentions *moderate chronic obstructive pulmonary disease*, is *moderate* (SNOMED 371927002) a qualifier entity apart from *chronic obstructive pulmonary disease* (SNOMED 13645005)? Or are they together a single entity (*Moderate chronic obstructive pulmonary disease [SNOMED 313297008]*)? What about the case of *untreated intraductal carcinoma in situ of unspecified breast?*

**Data table uncertainty** is caused by the CDM’s design of storing medical history in a separate location from active problems. As per Convention no. 4 in the definition of the CONDITION_OCCURRENCE table, “past diagnoses (‘history of’) are not recorded in the CONDITION_OCCURRENCE table. Instead, they are listed in the OBSERVATION table.” However, today’s active problems are tomorrow’s medical history, and from the perspective of observational, retrospective analyses, all patient data is historical. Furthermore, unless the data holders continuously update and move records between tables as needed, at the time of the query (phenotyping) it is unknown whether, at the time the data was recorded, any given piece of information was deemed active or historical.

**Conclusion**

This presentation calls for attention to semantic interoperability gaps in defining electronic phenotypes using a standardized data model. From our experience with specifying phenotypes for the eligibility criteria of 1,000 recent Phase 4 clinical trials, we found that these “real world” clinical researchers’ needs illuminate aspects that can be further discussed to allow for more robust secondary use of patient-level medical records.

**References**

Piloting Telemedicine for Emergency Treatment of Opioid Use Disorder

James R. Langabeer II, PhD1; Kimberly Chambers, MD2; Andrea Yatsco, PhD1; Tiffany Champagne-Langabeer, PhD1

1The University of Texas Health Science Center, School of Biomedical Informatics, Houston ER Opioid System, Houston, Texas; 2 The University of Texas Health Science Center, McGovern Medical School, Department of Emergency Medicine, Houston, Texas

Abstract

Opioid mortality rates continue to rise annually, but treatment is often inaccessible. Availability of emergency treatment involving partial agonist medication can be beneficial for initiating recovery; but geographic, transportation, and other barriers limit access. Telemedicine is one tool with potential to transform emergent interventions for individuals with opioid use disorder (OUD). In this clinical research study, we developed and piloted a telemedicine solution to examine feasibility, and here, we describe our study framework and initial outcomes.

Introduction

Telemedicine technology has had significant diffusion in recent years for both inpatient and outpatient care, but in addiction medicine it has not been widely adopted. This is due in part to requirements that the Drug Enforcement Agency (DEA) has for use of telemedicine and prescribing controlled substances. However, as mortality rates involving opioids continue to rise, as well as addiction rates overall, it is essential to identify novel ways to initiate and maintain treatment. Technology solutions are being proposed, including both mobile health and telemedicine; however, there have been few published manuscripts demonstrating feasibility of the conceptual framework. For individuals with OUD and those who have experienced non-fatal overdoses, it is critical to initiate rapid treatment while individuals are still experiencing withdrawal symptoms and before they re-engage in prior behaviors. Barriers to entering recovery programs exist for patients, including lack of familiarity with where and how to enter treatment, a scarcity of addiction medicine providers particularly in some geographic areas, cost of care, and limited transportation. Published national research suggests there are significant regional gaps in current treatment availability.1 Telemedicine could offer one solution.2 In this study, we present the framework for a telemedicine solution developed in Houston, Texas.

Objective

The objective of this research was to demonstrate the feasibility of a pilot telemedicine program for initiating medical treatment of patients with OUD and to analyze initial outcomes for treatment retention after the telemedicine intervention.

Methods

This study was conducted between August 2018 and February 2019 in Houston Texas. We designed an interventional, single-arm study. The process involved three steps: needs analysis and process flow (first three months); testing (fourth month); and deployment (final 3 months). A commercial, HIPAA-compliant telemedicine solution was implemented using both stationary equipment (in three locations) as well as a mobile solution for phones and tablets. The needs analysis and protocol involved establishing the originating site (location of the patient accompanied by an advanced nurse practitioners and a licensed social worker) and a distant site (involving DEA specially registered physicians). Due to telemedicine policy restrictions, significant effort went into establishing the correct infrastructure to support telemedicine. All providers were officially employees of the same university. The protocol consisted of obtaining initial enrollment information, baseline history data, and informed consent at an enrollment site. Advanced nurse practitioners (ANP) provided the medical screening and clinical assessments (e.g., clinical opiates withdrawal scale, diagnosis of OUD, confirmation of opioid in urine drug screen). This information was provided electronically to affiliated physicians, board certified in emergency medicine at a large urban hospital. The physicians were specially trained to provide addiction treatment and received a Drug Addiction Treatment Act (DATA 2000) waiver from the DEA, which is required to prescribe opioid agonist medication for treatment of OUD. Patients were seen through synchronous video conferencing. Physicians assessed the patients for opioid dependence, completed a brief history and medical exam, and ensured patients were in at least moderate withdrawal. Patients were followed longitudinally, through daily and weekly telephonic follow-ups to measure retention in treatment 30-days post-telemedicine
Inclusion criteria consisted of adults presenting with OUD or recent non-fatal overdose.

**Figure 1: Telemedicine Process Flow for Opioid Treatment**

Results

There were 10 patients presenting with OUD treated by telemedicine in the pilot study. Baseline characteristics for this sample are shown in Table 1. After 30 days, nine patients (90%) treated through the telemedicine encounter were still actively involved in medication-assisted treatment.

**Table 1: Patient Characteristics**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average age, y</td>
<td>35.1 (±12)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>6 (60%)</td>
</tr>
<tr>
<td>Female</td>
<td>4 (40%)</td>
</tr>
<tr>
<td>Race/Ethnicity:</td>
<td></td>
</tr>
<tr>
<td>Caucasian, non-Hispanic</td>
<td>6 (60%)</td>
</tr>
<tr>
<td>African American</td>
<td>3 (30%)</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>1 (10%)</td>
</tr>
<tr>
<td>Opioid type:</td>
<td></td>
</tr>
<tr>
<td>Intravenous heroin</td>
<td>7 (70%)</td>
</tr>
<tr>
<td>Pills</td>
<td>3 (30%)</td>
</tr>
<tr>
<td>Homeless, unstable housing</td>
<td>5 (50%)</td>
</tr>
</tbody>
</table>

Discussion

Feasibility of a model for utilizing telemedicine in the emergency initiation of opioid agonist treatment was demonstrated in this pilot project. Initial outcomes, 30 days after telemedicine encounter, are promising and suggest that immediate treatment of individuals remotely through telemedicine can positively initiate and retain individuals into treatment for the short-term, and help mitigate geographic, transportation, and cost barriers to care. There are policy implications for telemedicine opioid treatment which need to be explored, including federal policy restrictions for widespread use of telemedicine. We recommend that a larger randomized clinical trial be conducted over longer study periods further confirm the initial feasibility and efficacy from this pilot study.

References


A Model for Building an Analytics Based Decision Making Culture to Support Learning Healthcare Organizations

Cynthia LeRouge\textsuperscript{1}, George Sayre\textsuperscript{2,3}, Bianca Frogner\textsuperscript{2}, Paulo Gomes\textsuperscript{1}, Lisa V. Rubenstein\textsuperscript{4,5}, Savitha Sangameswaran\textsuperscript{2}

\textsuperscript{1}Florida International University, Miami, FL; \textsuperscript{2}University of Washington, Seattle, WA; \textsuperscript{3}VA HSR&D Center of Innovation (COIN) for Veteran Centered and Value Driven Care; \textsuperscript{4}University of California, Los Angeles, CA; \textsuperscript{5}RAND Corporation

Introduction

Many healthcare organizations integrate data analytics as part of their current efforts to improve patient access, satisfaction and quality of care. Effective data analytics requires a movement from data to information, to knowledge, to wisdom\textsuperscript{1} as proposed by the DIKW model. This progression towards a “data informed culture” requires introducing new systems, knowledge, and skills.

This study explores early frontline experiences in the Veterans Health Administration (VHA), under the Clinical Practice Management (CPM), an initiative driven by congressional mandate, which leverages data analytics to support overall monitoring and improving access to primary care across multiple levels and services. We use this context to study and model how healthcare organizations can cultivate a cultural transformation of “data to information to knowledge.” Such transformation efforts are critical for learning organizations to gain the insight needed to formulate intelligent responses to analytics, and the wisdom on how to leverage resources to actualize a response strategy.

Methods

This qualitative study used semi-structured interviews to collect data from 56 key informants associated with five VHA sites prototyping the CPM initiative to improve primary care access. Data was collected at multiple points in time over the first year of implementation. The interview protocols were guided by the Consolidated Framework for Implementation Research\textsuperscript{2} (CFIR) and included open-ended questions and structured prompts designed to elicit rich descriptions regarding: overall impressions, practice management strategies, goals, activities, relationship to clinic roles, data analytics usage, challenges, barriers, and facilitators. Key informants included national (central office) staff engaged with program implementation, Group Practice Managers (a new role created under the CPM), clinic leaders, and frontline staff. Inductive and deductive coding methods were used to identify key themes from interview transcripts. Data analysis was completed when the team reached consensus that all quotes were appropriately coded, and the resulting themes were stable.

Results

We distill our findings into a model for cultivating analytics based culture, as depicted in Figure 1.

Figure 1. Model for cultivating analytics based culture to support learning organizations.
Through analysis of empirical data, we identified five key action domains that align with the socio-technical nature of building a foundation of “data to information to knowledge”, as described in Table 1. Three domains relate to shoring up data analytics structures; and two domains are about mobilizing a CPM culture.

Table 1. Action domains

<table>
<thead>
<tr>
<th>Action Domains</th>
<th>Supporting Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SHORE UP DATA ANALYTIC STRUCTURES</strong></td>
<td></td>
</tr>
</tbody>
</table>
| Democratize the Data | • Enable access to data  
  • Release data from silos  
  • Move from data as a performance scorecard to objective gauge |
| Determine the Right Metrics | • Ensure covering all relevant outcomes (e.g., quality, access, patient satisfaction)  
  • Provide meaning to metrics through explicit definitions |
| Invest in the Right Tools | • Accommodate data summary & drill down (dashboards)  
  • Avoid Information and tool overload  
  • Organize and structure Information to help discern patterns  
  • Accommodate timely reporting |
| **MOBILIZE A CPM CULTURE** | |
| Invest in the Right People | • Formalize data analytical roles & leadership  
  • Cultivate a workforce willing to learn and adapt to analytic tools and processes  
  • Train people to identify actionable options in response to metrics (e.g. case based training) |
| Engage in analysis | • Promote metrics as core decision support tools  
  • Recognize counterbalancing metrics in decision making |

Discussion

This study demonstrates that a proactive, socio-technical approach that considers the path from data to wisdom is needed if we are to leverage increasingly massive sets of clinical and operational data collected in the modern healthcare space.

By developing a model that recognizes the importance of building analytics based decision making culture to support systems striving to be learning organizations, this study provides insights and approaches that can help inform the efforts of health care organization beginning to aggregate data and leverage analytics as part of their efforts to improve patient access, satisfaction, and quality of care.

References


Acknowledgments

This work was supported by the United States Department of Veterans Affairs Office of Veterans Access to Care (OVAC). The views expressed in this abstract are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs.
Measuring the Quality of Demographic Data Captured in Maryland’s Hospital and Emergency Department Discharges

Ashley Li, BS1, Hadi Kharrazi, MD PhD2
1Johns Hopkins University, Biomedical Engineering, Baltimore, MD; 2Johns Hopkins School of Public Health, Baltimore, MD

Introduction

Data quality remains an important issue for many sources of healthcare data such as hospital discharges, EHRs, and HIEs. Past studies have addressed completeness and accuracy of various health data sources; however, those studies focused on measuring the quality of clinical data and few examined demographic variables like age and race. In this study, the Maryland Healthcare Cost and Utilization Project’s (HCUP) inpatient, ambulatory surgery and services, and emergency department databases for the years 2013 and 2014 were examined for data quality issues. The data fields examined were sex, age, race, marital status, and zip code.

Methods

Maryland HCUP’s demographic data (2013 and 2014) were assessed for completeness, fidelity, and plausibility. Completeness ensures the lack of missing values, while fidelity questions the data’s dependability, and plausibility determines whether data is believable. These rules were derived from prior work in establishing a common data quality terminology. To ascertain these data qualities, the “visitlink” data element was used as the unique patient identifier across all HCUP’s underlying data sources: state inpatient database, state ambulatory surgery and services database, and the state emergency department database.

Completeness was measured by the proportion of non-missing data; the null count for a data element was compared with the patient’s total number of visits. Fidelity was assessed for sex, age, race, and marital status. Sex was assumed to not change due to the rarity of sex reassignment procedures. Age could not be less than 0, greater than 120, a range greater or equal to 2 within a year, a range greater or equal to 3 within 2 years, or standard deviation greater than 0.5. Patients who died in 2013 were flagged if reported to visit in 2014. Race was assumed to not change. Patients were flagged if they reported two or more non-null race statuses. In Maryland, it is illegal for minors younger than 15 to get married thus patients younger than 15 were flagged for non-individual or null marital statuses. Plausibility was assessed for marital status and zip code. For marital status, patients were flagged for 3+ marital status changes in a 60-day period for switching between individual and married, divorced and widowed, married and legally separated, or married and divorced. Zip code violations of plausibility included 4+ distinct zip codes in a year or 3+ zip code changes in a 60-day period.

After flagging patients for sex, age, race, marital status, or zip code data quality issues, a population specification table was created for the starting population, removed population, and remaining population. To compare the populations, a t-test was used for age, a binomial test for sex, and chi-square goodness of fit for marital status and race. The completeness, fidelity, and plausibility analysis were performed using R version 3.4.3 and MySQL 6.3 CE.

Results

After performing a completeness analysis, the starting population of 3,263,864 dropped to 3,076,908. 5.73% of the starting population was flagged for completeness errors. No patients were flagged for completeness issues in sex, but 245 were flagged for age, 146,274 for race, 37,523 for marital status, and 2,941 for zip code. Then plausibility and fidelity analyses were performed on the remaining patients leading a final population of 2,956,513 a drop of 9.42% from the starting population. After aggregating issues across demographic variables, 790 were flagged for age, 257,026 for race, 61,817 for marital status, and 5,254 for zip code (Table 1). Since sex had no issue in any aspect, it was dropped from the analysis. 307,351 patients had at least one issue among age, race, marital status, or zip code. 32 patients had an issue with age, race, marital status, and zip code simultaneously (Figure 1). The ‘cleaned’ population resembled the starting population’s age (mean and SD), sex, race and marital status distribution; however, all variables showed a statistically significant difference from the original population (p-value < .001).
Table 1. Completeness, fidelity, plausibility of demographic data in Maryland’s HCUP data

<table>
<thead>
<tr>
<th>Variables</th>
<th># Patients with Data Quality Issues</th>
<th>% of Variable Group</th>
<th>% of Total Excluded Population</th>
<th>% of Total Starting Population</th>
<th>Completeness, Fidelity, Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>790</td>
<td>100.00</td>
<td>0.28</td>
<td>0.02</td>
<td>C</td>
</tr>
<tr>
<td>• Missing</td>
<td>245</td>
<td>31.01</td>
<td>0.08</td>
<td>0.01</td>
<td>F</td>
</tr>
<tr>
<td>• Other</td>
<td>508</td>
<td>64.30</td>
<td>0.17</td>
<td>0.02</td>
<td>F</td>
</tr>
<tr>
<td>• Death rules</td>
<td>37</td>
<td>4.68</td>
<td>0.01</td>
<td>0.00</td>
<td>F</td>
</tr>
<tr>
<td>Race</td>
<td>257,026</td>
<td>100.00</td>
<td>83.62</td>
<td>4.48</td>
<td>C</td>
</tr>
<tr>
<td>• Missing</td>
<td>146,338</td>
<td>56.94</td>
<td>47.61</td>
<td>4.48</td>
<td>C</td>
</tr>
<tr>
<td>• Other</td>
<td>110,688</td>
<td>43.06</td>
<td>36.01</td>
<td>3.39</td>
<td>F</td>
</tr>
<tr>
<td>Marital Status*</td>
<td>61,817</td>
<td>100.00</td>
<td>20.11</td>
<td>1.89</td>
<td>C</td>
</tr>
<tr>
<td>• Missing</td>
<td>51,569</td>
<td>83.42</td>
<td>16.78</td>
<td>1.58</td>
<td>C</td>
</tr>
<tr>
<td>• 60-day rules</td>
<td>8,232</td>
<td>13.32</td>
<td>2.68</td>
<td>0.25</td>
<td>F</td>
</tr>
<tr>
<td>• Age</td>
<td>2,035</td>
<td>3.29</td>
<td>0.66</td>
<td>0.06</td>
<td>F</td>
</tr>
<tr>
<td>Zip Code**</td>
<td>5,254</td>
<td>100.00</td>
<td>1.71</td>
<td>0.10</td>
<td>F</td>
</tr>
<tr>
<td>• Missing</td>
<td>4,341</td>
<td>82.62</td>
<td>1.41</td>
<td>0.13</td>
<td>C</td>
</tr>
<tr>
<td>• Annual</td>
<td>360</td>
<td>6.85</td>
<td>0.12</td>
<td>0.01</td>
<td>P</td>
</tr>
<tr>
<td>• 60-day rules</td>
<td>711</td>
<td>13.53</td>
<td>0.23</td>
<td>0.02</td>
<td>P</td>
</tr>
<tr>
<td>Total Excluded Population</td>
<td>307,351</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion & Conclusion

Data quality can affect the analysis that informs decisions and policy. Literature did not address HCUP’s data quality specifically the temporality of demographic fields. We sought to show data quality issues in HCUP’s demographic data while addressing completeness, fidelity, and plausibility. Through this analysis, ~9.42% of the starting population was removed to form the final population. The removed population differed from the final population in race and marital status composition especially with the removal of null entries. Based on the rules, the removed population may encompass patients of multi-racial background because race comprises 83.62% of the removed population.

There are limitations on the study. Maryland’s HCUP database is highly curated. Several assumptions of temporal patterns and time limits were made but we tried to keep it logical and practical. Future research should examine the best data quality rules, temporal patterns, and time limits for hospital discharge data. Although atemporal issues remain the focus of data quality assessments, temporal patterns represent an area of further data quality research.

References

Detecting unanticipated actions downstream from clinical decision support: a data mining approach

Ron C. Li, MD, Imon Banerjee PhD, Daniel Rubin MD MS, Jonathan H. Chen MD PhD
Department of Biomedical Data Sciences, Stanford University, Stanford CA; Department of Medicine, Stanford University, Stanford CA; Department of Radiology, Stanford University;

Introduction
While there have been increasing efforts to develop more accurate algorithms for clinical decision support (CDS), less attention has been placed on understanding the downstream clinical decisions associated with CDS when deployed in the real world setting. Small scale ethnographic studies involving direct observations of clinicians using CDS are costly, yield limited insights, and are subject to the Hawthorne effect. Other studies evaluating CDS using electronic health record (EHR) data typically rely on measurements of a priori defined process or outcome metrics with limited insight into the unanticipated downstream actions taken by clinicians. For example, our institution, a quaternary care academic medical center, recently deployed a rules based early sepsis detection CDS that was initially evaluated with metrics such as frequency of appropriate antibiotics and diagnostic lab ordering after the CDS go live. These metrics, however, capture only a subset of possible downstream actions taken by clinicians after being exposed to the CDS in the real world setting, some of which are difficult to anticipate and may impact patient outcomes. Detection of unanticipated actions associated with CDS continue to rely on sporadic anecdotes and incident reports. This gap in our ability to systematically evaluate how CDS affects how clinicians think and behave hinders the design and implementation of safe and effective CDS for patient care. We demonstrate a data driven approach to detect downstream actions associated with a recently deployed severe sepsis CDS by mining ordering patterns in the EHR.

Methods
The severe sepsis CDS system is a rules-based alert that activates if a set of clinical criteria are met based on clinical data in the EHR. The alert is delivered via a text-page sent directly to the treating physician, and has been live on the inpatient medicine ward at our institution since 2014. To identify downstream clinician actions, we started by selecting all medication orders (mapped to RxNorm) entered for a given patient 24 hours after the firing of a sepsis alert. We combined sequential pattern mining and interactive graph representation to allow easy exploring of the CDS ordering space. First, we formulated the sequences (<A><B><C, D>) by maintaining temporal ordering pattern from the list of medication orders grouped by patient and alert where medication <A> was ordered before <B>, and <C> and <D> were ordered at the same time. Only one alert per 24 hour time period was considered for each patient and sequences were excluded if the first medication was ordered after one hour of alert firing. Afterwards, we applied a modified version of sequential pattern mining algorithm PrefixSpan (Prefix-projected Sequential pattern mining) which extracts the frequent sequential patterns through the use of “divide-conquer” method and also allows to compute the confidence of the sequential patterns. The method divides the raw sequence database into smaller projection databases, and then mine the sequential pattern in smaller databases given a predefined minimum support threshold. In this study, in order to capture both anticipated and unanticipated order pattern, we used a more flexible minimum support value which is above 5% of the total number of sequence.

To support interactive exploratory analysis of the discovered patterns, we computed a directed graph representation of the sequential pattern content where nodes are medication orders explored by modified PrefixSpan and directed connections represent ordering sequence. For example, {<A>+<B>} and {<A>+<C>} are two valid patterns, the graph (G) will have nodes <A>, <B>, <C> with a edges connecting the nodes <A,B> and <A,C>. The generated graph is interactive which allows the end-user to click on a
edge and visualize the support and confidence value for each transition of medications calculated by the sequence mining algorithm.

**Results**

A total of 6992 alerts fired for 4024 patients hospitalized on the inpatient medicine ward between 2014 and 2017. 5300 unique alerts from medicine were included in our analysis and 1425 sequences above the minimum support threshold were identified. Most sequences included items that are expected to be administered after the sepsis alert, such as an intravenous (IV) fluid bolus and vancomycin (Table 1). However, we also identified a frequent sequence containing furosemide, which was unanticipated, since administering a diuretic would be counter to the expected action taken for a patient with severe sepsis. Further, furosemide disconnected from other items in any other frequent sequences (Figure 1).

**Discussion**

We describe a novel approach to detect unanticipated actions downstream from a severe sepsis alert on an inpatient medicine ward. While expected actions such as administration of IV fluids and antibiotics were indeed identified as common sequences, we did discover an unanticipated frequent sequence consisting of furosemide administration. Furosemide is a diuretic, which is atypical and in fact often opposite of what is expected to be given for patients with severe sepsis. We also see from the graph that furosemide exists separately from the other more connected items in frequent sequences, which indicate that this may represent a distinct decision pathway. It is possible that the furosemide containing sequences represent patients who triggered false positive alerts, or patients with both volume overload and severe sepsis. Further exploration of these previously unidentified insights may be important for iteration of CDS design and deployment strategy.

**Figures**

![Sequence](image.png)

**Figure 1**

**Figure 2**

**References**


Development and Evaluation of a Machine Learning-based Approach to Detect Errors in Pediatric Weight Data

Lei Liu, BS\textsuperscript{1,2}, P.J. Van Camp, MD\textsuperscript{1,2}, C. Monifa Mahdi\textsuperscript{1}, S. Andrew Spooner, MD\textsuperscript{1,2}, Danny T.Y. Wu, PhD, MSI\textsuperscript{1,2}, Yizhao Ni, PhD\textsuperscript{1,2}

\textsuperscript{1}Cincinnati Children’s Hospital Medical Center; \textsuperscript{2}University of Cincinnati, Cincinnati, OH, USA

Introduction

Inaccurate weight measures can cause critical safety events in pediatrics due to the frequent use of weight-based dosing. Previous efforts to mitigate weight errors, including interruptive alerts during data entry, rely heavily on manual vigilance and detect only a fraction of errors. Sustaining effective and accurate detection of weight errors remains a challenging issue among pediatric healthcare institutions. This study focused on developing a machine learning (ML)-based approach to analyze pediatric growth data and detect potential weight errors.

Methods

We collected de-identified patient weight data from Cincinnati Children’s Hospital Medical Center between 2010 and 2018. A total of 4.3 million weight points were collected from the institutional electronic health record system for 347,056 patients. Following our earlier study\textsuperscript{1}, weight points documented in the first 24 months were excluded because newborns can have more complex growth patterns. A subset of 4,000 patients with over 50,000 weight points were manually reviewed by three domain experts to create a reference standard set of weight errors (Cohen’s Kappa=0.628 and prevalence-adjusted and bias-adjusted Kappa=0.994)\textsuperscript{1}. For the full dataset, three established weight error detection methods, including a modified-z-score-based method developed by the Centers for Disease Control and Prevention (CDC)\textsuperscript{2}, a computerized approach for determining implausible weight values developed by Children’s Hospital of Philadelphia\textsuperscript{3}, and a linear regression method developed in our earlier study were applied to identify abnormal weight values. The majority vote of the three methods served as a pseudo reference set of weight errors\textsuperscript{4}.

The sub and full datasets were used to train and evaluate the predictive models. To assure that each chart had sufficient data to demonstrate its patterns and dynamics, patient charts with fewer than 4 weight points or with clear evidence of errors (e.g., weight z-score over 100) were excluded from the two datasets. For each data point we extracted seven variables to capture weight characteristics and growth dynamics: 1) subject weight in kilograms, 2) LMS z-score according to the subject’s sex and age\textsuperscript{2}, 3) the distance of the weight value from the reference median in units of one half of the distance between 0 and +2 z-scores\textsuperscript{2}, 4) percentage of the reference population that is below the weight\textsuperscript{2}, 5) absolute age difference from the previous weight point, 6) absolute weights difference from the previous weight point, and 7) absolute z-scores difference from the previous weight point. ML algorithms were developed to predict weight errors, including multivariate logistic regression with L2 normalization (LR), support vector machines with polynomial kernels (SVM-P), decision trees (DT), random forests (RF), and artificial neural networks (ANN).

We performed a stratified random sampling based on individual patients to split each dataset into two parts, 70\% for training and 30\% for evaluation. Ten-fold cross-validation was utilized in training the computerized algorithms, where model parameters were optimized with grid search parameterization. Stratified down-sampling was integrated into cross-validation to improve class imbalance. For model evaluation, we adopted the area under the receiver operating characteristic curve (AUC) as the primary measure and reported positive predictive value (PPV), sensitivity (SEN), negative predictive value (NPV), specificity (SPEC) when SEN reached 90\% (a level required for production). This study was supported by the Mary K Logan research award received by Dr. Wu in 2018.

Results

There were 3,692 patients with 45,892 weight points in the reference standard set and 222,508 patients with 3,009,762 in the full dataset. The error rates were 0.51\% (233 weight errors in 194 patients charts) and 0.14\% (4,278 weight errors in 3,360 patients charts) respectively. The model performances on the two datasets are presented in Table 1 and Table 2. The ML algorithms performed generally better on the full dataset with pseudo reference. RF achieved the best performance on both datasets (AUCs of 93.2%/99.8\% on the test data) although its improvements over LR and ANN were not significant at 0.05 level. Figure 1 visualizes four example charts representing true positive (TP), false positive (FP), all true negative (TN) and false negative (FN) predictions of the best performing RF on the reference standard test set. To understand rules determining a weight error, Figure 2 presents the best-performing decision tree (AUC of 87.2\%) developed on the reference standard set.

Discussion and Conclusion

The ML approach demonstrated good performance in detecting abnormal weights in pediatric weight data. The PPV and SEN achieved by the best-performing RF on the reference standard set suggested that with using the ML approach a clinician could capture 90\% of errors by reviewing approximately 20\% of weight points. The TP/TN examples...
suggested that RF was good at learning pediatric growth trends to capture sudden weight changes. However, it would fail when there was paucity of data (the FP/FN examples), revealing a major limitation of ML-based methods. The analysis of DT suggested that data points with higher z-score and weight changes within a short time period (as evidenced by age difference) have higher risk of being errors. Finally, all algorithms showed better performance on the full dataset because the pseudo reference was more determinate than human judgment. As such, our future work involves annotating the full dataset to further evaluate the algorithms in a more realistic setting.

By visualizing ML predictions on weight data, timely correction could be made to avoid weight errors and to mitigate its effects on patient safety. Consequently, we hypothesize that the ML-based approach, when rolled out for production, holds great potential to significantly reduce medication safety events among pediatric patients.

Table 1. Performance of different machine learning algorithms on the reference standard set.

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>PPV</th>
<th>SEN</th>
<th>NPV</th>
<th>SPEC</th>
<th>AUC</th>
<th>PPV</th>
<th>SEN</th>
<th>NPV</th>
<th>SPEC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>1.8%</td>
<td>93.6%</td>
<td>99.8%</td>
<td>47.6%</td>
<td>88.2%</td>
<td>0.7%</td>
<td>90.5%</td>
<td>99.9%</td>
<td>37.0%</td>
<td>86.7%</td>
</tr>
<tr>
<td>DT</td>
<td>0.5%</td>
<td>99.0%</td>
<td>75.0%</td>
<td>0.01%</td>
<td>81.7%</td>
<td>0.5%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0.4%</td>
<td>87.2%</td>
</tr>
<tr>
<td>RF</td>
<td>2.5%</td>
<td>94.3%</td>
<td>100.0%</td>
<td>64.2%</td>
<td>92.2%</td>
<td>2.0%</td>
<td>90.5%</td>
<td>100.0%</td>
<td>78.9%</td>
<td>93.2%</td>
</tr>
<tr>
<td>SVM-P</td>
<td>1.3%</td>
<td>93.6%</td>
<td>100.0%</td>
<td>52.3%</td>
<td>84.3%</td>
<td>1.1%</td>
<td>90.5%</td>
<td>100.0%</td>
<td>63.2%</td>
<td>89.4%</td>
</tr>
<tr>
<td>ANN</td>
<td>3.4%</td>
<td>93.6%</td>
<td>100.0%</td>
<td>55.4%</td>
<td>90.4%</td>
<td>0.8%</td>
<td>90.5%</td>
<td>100.0%</td>
<td>48.8%</td>
<td>90.1%</td>
</tr>
</tbody>
</table>

Table 2. Performance of different machine learning algorithms on the full dataset.

<table>
<thead>
<tr>
<th>ML Algorithm</th>
<th>PPV</th>
<th>SEN</th>
<th>NPV</th>
<th>SPEC</th>
<th>AUC</th>
<th>PPV</th>
<th>SEN</th>
<th>NPV</th>
<th>SPEC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>22.6%</td>
<td>90.2%</td>
<td>100.0%</td>
<td>99.5%</td>
<td>99.6%</td>
<td>26.3%</td>
<td>90.0%</td>
<td>100.0%</td>
<td>99.7%</td>
<td>99.5%</td>
</tr>
<tr>
<td>DT</td>
<td>53.2%</td>
<td>91.3%</td>
<td>100.0%</td>
<td>99.4%</td>
<td>97.8%</td>
<td>52.1%</td>
<td>91.2%</td>
<td>100.0%</td>
<td>99.9%</td>
<td>99.5%</td>
</tr>
<tr>
<td>RF</td>
<td>57.4%</td>
<td>90.2%</td>
<td>100.0%</td>
<td>99.9%</td>
<td>99.8%</td>
<td>57.5%</td>
<td>90.0%</td>
<td>100.0%</td>
<td>99.9%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM-P</td>
<td>1.0%</td>
<td>90.2%</td>
<td>100.0%</td>
<td>86.2%</td>
<td>92.9%</td>
<td>9.57%</td>
<td>90.0%</td>
<td>100.0%</td>
<td>98.9%</td>
<td>98.8%</td>
</tr>
<tr>
<td>ANN</td>
<td>34.0%</td>
<td>90.2%</td>
<td>100.0%</td>
<td>97.2%</td>
<td>99.7%</td>
<td>35.4%</td>
<td>90.0%</td>
<td>100.0%</td>
<td>99.8%</td>
<td>99.6%</td>
</tr>
</tbody>
</table>

Figure 1. Example predictions using random forest.

Figure 2. The best-performing decision tree.

References
YouTube Video Analytics for Patient Self-Care of Chronic Diseases

Xiao Liu, PhD¹, Bin Zhang, PhD², Anjana Susarla, PhD³, Rema Padman, PhD⁴
¹University of Utah, Salt Lake City, UT, USA, ²University of Arizona, Tucson, AZ, USA, ³Michigan State University, Lansing, MI, USA, ⁴Carnegie Mellon University, Pittsburgh, PA, USA

Introduction

Chronic diseases, such as heart disease, cancer and diabetes, are among the most expensive and yet preventable health problems (WHO 2013). Effectively managing chronic conditions through patient self-care is a major focus of health policy. However, self-care is a complex set of activities and requires significant amount of medical knowledge. Only 12% of the US population are estimated to be proficient in health literacy (HHS 2008). Most people face difficulties in interpreting and using health information. Purely text-based medical instructions result in poor patient attention, comprehension, recall, and adherence, which translates to poorer health outcomes and higher healthcare costs in the overall economy, especially for patients with low literacy levels (Moorehead et al. 2013).

Visual social media sites such as YouTube host videos containing information on the pathogenesis, diagnosis, treatments, and prevention of various conditions. These health-related videos offer an effective way to understand medical knowledge and to bridge this health literacy gap (Madathil et al. 2014). Nevertheless, healthcare videos on YouTube range from homemade ones expressing personal opinions and experiences to videos made by reputable healthcare entities and expert clinicians for professional education (Pant et al. 2012). Due to the heterogeneity of content quality and content helpfulness (Pant et al. 2012) of visual social media sites, patients, healthcare providers and government agencies have expressed concerns about the quality and reliability of such information. In this study, we examine the impact of encoded medical information on user engagement by proposing an interdisciplinary lens that synthesizes deep learning methods to understand the level of medical information encoded in YouTube videos and its impact on the distinct types of user engagement.

Methods

We present our research approach in Figure 1. We collected YouTube videos related to diabetes care with the YouTube Data API and 200 diabetes related search keywords. We fetched the top 100 videos according to the YouTube’s search ranking for each search term and store the ranking of returned videos and their metadata in a database for further analysis. The information we collected includes channel ID (account name), publish time of the video, video title, video description, video tags, video duration, video definition, video caption availability, video rating, view count, like count, dislike count, and comment count. In total, we collect 19,873 unique videos.

We extracted medical terminology from the videos’ metadata using a deep learning method. We devised a BLSTM model to extract medical terms from the user-generated video descriptions at the sentence level. The video descriptions were first tokenized into individual sentences and words. The BLSTM model achieved 95% in precision, 91.8% in recall, and 92.9% in F1 for medical term extraction. A machine-learning model was devised to classify the encoded medical information in videos using the inputs from deep learning and video level features. These two research associates independently viewed these 600 videos and ultimately labeled the videos for high or low medical information. The high or low medical information categorization was based on the annotators’ experience as a consumer/viewer of the videos. A domain expert (a medical doctor) viewed all videos and consolidated the annotation results. The inter-rater reliability for classifying encoded medical information in videos is 0.92.

Figure 1. Proposed Research Approach

We present our research approach in Figure 1. We collected YouTube videos related to diabetes care with the YouTube Data API and 200 diabetes related search keywords. We fetched the top 100 videos according to the YouTube’s search ranking for each search term and store the ranking of returned videos and their metadata in a database for further analysis. The information we collected includes channel ID (account name), publish time of the video, video title, video description, video tags, video duration, video definition, video caption availability, video rating, view count, like count, dislike count, and comment count. In total, we collect 19,873 unique videos.

We extracted medical terminology from the videos’ metadata using a deep learning method. We devised a BLSTM model to extract medical terms from the user-generated video descriptions at the sentence level. The video descriptions were first tokenized into individual sentences and words. The BLSTM model achieved 95% in precision, 91.8% in recall, and 92.9% in F1 for medical term extraction. A machine-learning model was devised to classify the encoded medical information in videos using the inputs from deep learning and video level features. These two research associates independently viewed these 600 videos and ultimately labeled the videos for high or low medical information. The high or low medical information categorization was based on the annotators’ experience as a consumer/viewer of the videos. A domain expert (a medical doctor) viewed all videos and consolidated the annotation results. The inter-rater reliability for classifying encoded medical information in videos is 0.92.
Table 1. Medical Information Classification Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Medical Information Videos</td>
<td>88.7%</td>
<td>83.9%</td>
<td>86.2%</td>
</tr>
<tr>
<td>Low Medical Information Videos</td>
<td>86.6%</td>
<td>90.6%</td>
<td>88.6%</td>
</tr>
<tr>
<td><strong>Overall Accuracy</strong></td>
<td><strong>87.5%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The performance of the classification is reported in Table 1. We deployed this classification model to classify the remaining videos in our collection. In total, we have 7,948 videos classified as high medical information and 11,325 videos classified as low medical information.

**Results and Discussions**

PCA reveals three broad categories of user engagement: non-engagement, sustained attention driven engagement, and selective attention driven engagement. Sustained attention driven engagement relates to measures of views, comments, likes, and dislikes. Selective attention driven engagement strongly relates to measures of relevance, which reveals the extent to which the content of a video matches the interests of users. Table 2 illustrates the estimated treatment effect on the three principal components of collective engagement.

Table 2. Estimated Treatment Effect

<table>
<thead>
<tr>
<th>Component</th>
<th>Estimate</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-engagement (Intercept)</td>
<td>0.78</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Average Treatment Effect</td>
<td>-2.98</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Sustained attention (Intercept)</td>
<td>0.48</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Average Treatment Effect</td>
<td>-1.54</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Selective attention (Intercept)</td>
<td>-0.05</td>
<td>0.18</td>
</tr>
<tr>
<td>Average Treatment Effect</td>
<td>0.03</td>
<td>0.09</td>
</tr>
</tbody>
</table>

We find that medical information is significantly negatively linked to the measure of non-engagement. It implies that videos containing valuable medical information are more engaging to a user. Secondly, we find a positive and marginally significant relationship between medical information encoded in the video and selective attention driven engagement. This suggests that high medical information can increase the chance for the initiation of engagement. At the same time, we also find a negative relationship between medical information and sustained attention driven engagement. The higher density of medical information does not necessarily result in sustained attention driven engagement. Videos with a higher amount of encoded medical information may lead to disengaged users. The complexity of encoded medical information may discourage users from deeper interaction due to the high cognitive load to comprehend the information.

**Conclusion**

Our study underscores that it is crucial for clinicians and practitioners to understand how patients engage with medical information in social media settings. The Internet has reduced much of the information asymmetry between healthcare practitioners and consumers by providing multiple avenues whereby patients can educate themselves with user generated content. While healthcare organizations or consumers produce patient educational materials, not only should they think about what the medical information to deliver but also how to meet the interest and health literacy level of the consumers.

**References**

Analyzing Flowsheet Data to Refine and Validate Information Models

Kay S. Lytle, DNP, RN-BC, NEA-BC, CPHIMS, FHMIS1, Bonnie L. Westra PhD, RN, FAAN, FACMI2, Mary Hook, PhD, RN-BC3, Mischa Adams, MSN, RN, CCRN4, Samira Ali, DNP, RN5, Meg Furukawa, MN, RN-BC6, Steven G. Johnson, PhD2, Tari Rajchel DNP, RN7, Sarah Collins Rossetti, RN, PhD, FACMI, FAMIA8, Theresa (Tess) Settergren, MHA, MA, RN-BC9, Luann Whittenburg, PhD, RN-BC, FNP, FAAN10

1Duke University Health System, Durham, North Carolina; 2School of Nursing, University of Minnesota, Minneapolis, Minnesota; 3AdvocateAuroraHealth, Milwaukee, Wisconsin; 4Allina Health, Minneapolis, Minnesota; 5School of Nursing, Grand Canyon University, Phoenix, Arizona; 6University of California Los Angeles Health, Los Angeles, California; 7North Memorial Medical Center, Robbinsdale, Minnesota; 8Columbia University, New York, New York; 9Retired; 10Bumrungrad International Health Informatics/Health System Architecture, Bangkok, Thailand

Introduction
Flowsheets provide a rich source of clinical data and represent a major source of documentation within the electronic health record (EHR); however, the content is not standardized within and across organizations. While the expansion of EHRs provides organizations the opportunity to evaluate, share and compare data for quality improvement and research strategies to improve the effectiveness, efficiency and care outcomes, this lack of standardization presents a significant barrier. “Information models” represent clinical concepts, definitions, and their relationships for consistent documentation and sharing across EHR platforms1. Information models (IMs) with associated data standards normalize differences in documentation to improve data sharing. Nursing data are underrepresented in much IM work. Westra et al.2 used flowsheet data from one health system with over 2.4 million patients to develop 10 reference IMs. The current work represents a nation-wide effort to bring leaders from large, integrated health systems to conduct a “bottoms up” approaches to validate similarities and identify differences, gain consensus, and inform future data standards and EHR design.

This study used a data-driven approach to validate and refine existing IMs with metadata extracted from the EHRs of multiple large integrated health systems to refine and extend the models with value sets. Three IMs were validated: pain; genitourinary (GU) system including catheter associated urinary tract infections (CAUTI); and falls. Pain represents a commonly occurring problem assessed and managed by all nurses, and CAUTI and falls are nurse-sensitive measures3. Flowsheet metadata were analyzed to standardize concepts, definitions, and associated values sets for assessments, goals, interventions and outcomes, recognizing that some content is not captured in flowsheets. The goal was to provide direction for improving the ability to share and compare nursing data and improve patient outcomes in the future.

Methods
A retrospective observational study was conducted using an iterative consensus-based approach to map, analyze, and evaluate data contributed by nursing informatics researchers from 10 health care organizations. A convenience sample of participants active in the Nursing Knowledge Big Data Science Initiative represented their organizations as participants in the study4. Organizations that provided flowsheet data included Allina Health, Aurora Health Care, Cedars Sinai, Duke University Health System, Fairview Health Services, Kaiser Permanente, North Memorial Medical Center and UCLA Health. Organizations that provided information about their EHR content included Bumrungrad International Hospital and Partners Healthcare (for pain IM only). Each organization determined if IRB approval was required and if so obtained it to participate in the study.

De-identified metadata representing the actual flowsheets used in patient care were extracted from the clinical databases of participating organizations, securely transferred, and imported into a secure server at the University of Minnesota. Meta-data included a unique identifier for each flowsheet row, internal description and name for the flowsheet row (concept/question), name of the groups and templates from data entry screens, the pick list values (answers) associated with the flowsheet row, the number of observations, encounters and patients, and the data of first and last uses. FloMap data mapping software supported aggregation and comparison of data across organizations5. Each informatics specialist mapped their organizational metadata to the established concepts in the pain, GU and falls reference IMs using FloMap. The mappings aided in identifying similarities and differences

1397
within and across organizations to inform the pain, GU and falls IMs and terminology. Reports were generated for comparison across organizations to drive consensus during bi-weekly meetings. Consensus addressed which concepts to keep, combine, remove, or add. The group discussed use cases, evidence-based practice, regulatory and accreditation requirements to inform consensus, develop definitions and determine the associated value sets.

**Results**

The aggregated metadata from the EHRs of eight large health care organizations and the design content from two additional organizations represented flowsheet data from 6.6 million patients, 27 million encounters, and 683 million observations. A comparison of the reference and validated models for pain, GU and falls and their associated classes, concepts and value sets are represented below (Table 1). The validation examined the generalizability of the reference IM and used consensus to recommend the essential content with the removal or items that were redundant or rarely used.

**Table 1.** Classes and concepts for reference and validated information models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Pain</th>
<th>Genitourinary</th>
<th>Falls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Reference</td>
<td>Validated</td>
<td>Reference</td>
</tr>
<tr>
<td>Classes</td>
<td>12</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Concepts</td>
<td>-</td>
<td>30</td>
<td>38</td>
</tr>
<tr>
<td>Value Set Items</td>
<td>-</td>
<td>396</td>
<td>-</td>
</tr>
</tbody>
</table>

In most cases, the validated IM reduced the number of classes and concepts, such as the pain IM that contains 30 concepts, 4 panels (classes), and 396 value set items with descriptions, definitions and value sets for the most important concepts currently captured in nursing-based flowsheet documentation for pain management. The final IMs were more robust with the inclusion of goals and age/condition specific pain scales with a considerable reduction in the number of classes and concepts going forward. Genitourinary included a key concept for a within defined limits assessment and then classes for voiding assessment, urinary incontinence assessment, urine assessment genitalia assessment, genitourinary (GU) device type, GU device assessment, and GU interventions. The validated fall prevention information model expanded to seven classes while reducing to 43 concepts with 179 value set items and increasing the fall risk screening tools from two to 10. Overall the final IMs were more comprehensive, inclusive of more tools and scales, and streamlined for efficiency as appropriate. Examples of the models will be shared.

**Discussion**

The consensus-based IMs for pain, GU and falls represent refinements proposed after in-depth analysis of the actual meta-data generated by nurses in EHR flowsheet documentation from participating systems. The findings provide direction for improving the ability to share and compare nursing data and support strategies that will lead to improved patient outcomes in the future. The concepts and value sets are in the process of encoding with LOINC and SNOMED-CT. The encoded standards then will be available to extend common data models used to support nursing data exchange and research nationally and internationally.

**References**

Comparing the Distribution and Predictive Power of Medication Adherence Indexes Derived from EHRs and Claims in Forecasting Healthcare Utilization

Xiaomeng Ma, MS\textsuperscript{1}, Changmi Jung PhD\textsuperscript{2}, Hadi Kharrazi, MD PhD\textsuperscript{1}

\textsuperscript{1}Center for Population Health IT, Johns Hopkins School of Public Health, Baltimore, MD; \textsuperscript{2} Carey School of Business, Johns Hopkins University, Baltimore MD

Introduction

Electronic health records (EHRs) and administrative claims data capture medication (Rx) orders and fills. Previous studies have focused on extracting Rx information from a single data source (e.g., measuring Rx possession ratio). No attempt has been made to probe the correlation of Rx adherence indexes in linked EHRs and claims. Rx complexity, as reflected in EHRs, can impair patient Rx filling adherence recorded in claims, and may associate with adverse outcomes such as hospitalization and increased cost\textsuperscript{1}. We sought to explore constructing patient-level Rx adherence measures - Medication Regimen Complexity Index (MRCI) \textsuperscript{1}, Medication Possession Ratio (MPR) \textsuperscript{1}, and Rx Fill-rates\textsuperscript{2} using administrative data. By quantifying their correlation in a population, we can understand to which level medication complexity affects patients adhering behaviors. Subsequent models were also developed to improve utilization prediction.

Methods

The study population comprised 65,653 patients with continuous enrolment in a single insurance plan from 2012 to 2013\textsuperscript{2}. Patients had at least one outpatient visit within two years and did not reach 65 by the time of inclusion. We developed and compared MRCI, MPR and Fill-rates based on linked patient records extracted from the integrated single-provider system. MRCI is an instrument to quantify medication complexity for outpatients, synthesizing the complexity caused by dosing frequency, dosing form and additional directions on EHR e-prescribing. MPR is another common index, measuring the ratio of medication supplied days and intervals between fills on claims. Fill rates is an Rx indicator we have proposed previously\textsuperscript{2}, to evaluate medication adherence by capturing the prescription filled at the pharmacy within a designated interval (7 and 30 days). All Rx adherence indexes were constructed based on National Drug Codes (NDCs) in EHRs and claims. To accommodate condition like Rx substitutions at pharmacy counters, we mapped Rx NDCs to 62 RxMG prescribed medication defined morbidity groups (derived from Johns Hopkins Adjusted Clinical Groups® (ACG®)\textsuperscript{3}) to construct Fill rates. Outpatient prescription count and Charlson Comorbidity Index were also calculated at the patient level. Patients were stratified by their demographics and utilization history, to demonstrate the distribution of Rx indicators in subpopulations. Statistical analysis was performed to evaluate the correlation of MRCI and MPR.

To evaluate the impact of MRCI, MPR and Fill rates on prospective utilization predictions, we split the dataset by year, predicting 2013 outcomes based on 2012 data. Three base models were tested. Binary hospitalization indicators and total medical cost (capped at top 1% total cost) were outcomes of interest, and were analyzed with logistic regression and linear regression, respectively. We added MRCI, MPR and Fill rates respectively to the base models. Changes of interpretable variation in enhanced models in contrast to base models were examined to show the added values to prediction of these three medication indices. For validation, we conducted bootstrapping (200), splitting the complete dataset into 1:1 training and validation sets for performance measures of adjusted $R^2$ and AUC. Improved Integrated Discrimination Improvement (IDI)\textsuperscript{4} and $R^2$ were adopted to account for the significance of variation change. R v3.5.1 was used for statistical analysis.

Results

The mean age of 65,653 patients was 38±18 years old. Among sampled population, 13,854 (21.1%) and 4,630 (7.1%) patients had visited emergency department or hospitalized during 2012–2013. The mean prescription count and mean MRCI were 5.7±6.8 and 16.5±21.2 after dropping the inpatient prescriptions. The average of MPR for both years was 0.56±0.51, and 20,794 (42.8%) patients showed a very low level of medication adherence (MPR<0.5) to their outpatient prescriptions. Statistical analysis indicates MRCI and MPR were negatively correlated in a given year (Pearson correlation = -0.17). In comparison to MPR, the 30-day Fill rates and 7-day Fill rates in this population were 0.77±0.24 and 0.67±0.29. The flagged patient who hospitalized in any of these two years had more complex medical regimens (MRCI: 30.2±38.1 vs. 11.4±11.8), but low adherence to outpatient prescriptions (MPR: 0.49±0.30 vs. 0.58±0.30; 30-day Fill rates: 0.76±0.20 vs. 0.79±0.30) compared to outpatients (Table 1). A cohort of 48,435 patients had complete MRCI, MPR and Fill rates scores for predictive modeling. MRCI added predictive value to demographic base-models for hospitalization and medical cost predictions (IDI: 234.4 [198.0 – 273.2%]; improved $R^2$: 4.8% [4.2 – 5.2%]). However, the added-value decreased
when adding Charlson Index to the base model (IDI: 49.9 [38.3 – 63.4%]; improved \( R^2 \): 2.3% [2.2 – 2.4]). A similar pattern was observed for MPR, but with minimal predictive effect than MRCI. The added-value for hospitalization and total medical cost predictions were significantly decreased after adding ACG risk scores to the base model (Table 2).

### Table 1. Rx derived Indicators stratified by age, sex, hospitalization and chronic conditions

<table>
<thead>
<tr>
<th>Models</th>
<th>AUC (95% CI)</th>
<th>IDI (%)</th>
<th>( R^2 ) (%, 95% CI)</th>
<th>( R^2 ) Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base model 1: age + sex</td>
<td>0.6 (0.58, 0.61)</td>
<td>1.36 (1.13, 1.58)</td>
<td>0.00 (0.00, 0.00)</td>
<td></td>
</tr>
<tr>
<td><strong>Base model 1 + Fill 30days</strong></td>
<td>0.6 (0.59, 0.61)</td>
<td>234.4 (198, 273.2)</td>
<td>1.35 (1.13, 1.55)</td>
<td>-0.01 (0.00, 0.00)</td>
</tr>
<tr>
<td><strong>Base model 1 + MRCI</strong></td>
<td>0.64 (0.63, 0.66)</td>
<td>100.2 (94.4, 105.6)</td>
<td>6.13 (5.3, 6.82)</td>
<td>4.77 (4.16, 5.24)</td>
</tr>
<tr>
<td><strong>Base model 1 + MPR</strong></td>
<td>0.63 (0.62, 0.65)</td>
<td>246.6 (210.7, 283.7)</td>
<td>2.51 (2.05, 2.81)</td>
<td>1.14 (0.91, 1.23)</td>
</tr>
<tr>
<td><strong>Base model 2: age + sex + Charlson</strong></td>
<td>0.64 (0.62, 0.65)</td>
<td>7.9 (7.09, 8.63)</td>
<td>0.00 (0.00, 0.00)</td>
<td></td>
</tr>
<tr>
<td><strong>Base model 2 + Fill 30days</strong></td>
<td>0.64 (0.62, 0.65)</td>
<td>234.4 (198, 273.2)</td>
<td>1.35 (1.13, 1.55)</td>
<td>-0.01 (0.00, 0.00)</td>
</tr>
<tr>
<td><strong>Base model 2 + MRCI</strong></td>
<td>0.66 (0.64, 0.67)</td>
<td>100.2 (94.4, 105.6)</td>
<td>6.13 (5.3, 6.82)</td>
<td>4.77 (4.16, 5.24)</td>
</tr>
<tr>
<td><strong>Base model 2 + MPR</strong></td>
<td>0.66 (0.64, 0.67)</td>
<td>246.6 (210.7, 283.7)</td>
<td>2.51 (2.05, 2.81)</td>
<td>1.14 (0.91, 1.23)</td>
</tr>
<tr>
<td><strong>Base model 3: age + sex + ACG</strong></td>
<td>0.70 (0.68, 0.71)</td>
<td>17.8 (16.64, 19.06)</td>
<td>0.00 (0.00, 0.00)</td>
<td></td>
</tr>
<tr>
<td><strong>Base model 3 + Fill 30days</strong></td>
<td>0.70 (0.68, 0.71)</td>
<td>234.4 (198, 273.2)</td>
<td>1.35 (1.13, 1.55)</td>
<td>-0.01 (0.00, 0.00)</td>
</tr>
<tr>
<td><strong>Base model 3 + MRCI</strong></td>
<td>0.69 (0.68, 0.71)</td>
<td>100.2 (94.4, 105.6)</td>
<td>6.13 (5.3, 6.82)</td>
<td>4.77 (4.16, 5.24)</td>
</tr>
<tr>
<td><strong>Base model 3 + MPR</strong></td>
<td>0.70 (0.69, 0.72)</td>
<td>246.6 (210.7, 283.7)</td>
<td>2.51 (2.05, 2.81)</td>
<td>1.14 (0.91, 1.23)</td>
</tr>
</tbody>
</table>

### Table 2. Performance of enhanced models added with Rx derived indicators in contrast to base models

### Conclusion

This study primarily investigated the quantitative measures to evaluate patient medication adherence. High medication regimen complexity is associated with the low Rx filled at pharmacy among outpatients. Although patients tend to fill the their very first prescriptions on schedule, the overall adherence rate is low. The MRCI and MPR modestly improved prospective utilization predictions, although the effects are not significant for models using the ACG risk scores.

### References

Clinic Characteristics Associated with Screening for Social Determinants of Health

Miguel Marino, PhD1, Jorge Kaufmann, MS1, David Ezekiel-Herrera, MS1, Dagan Wright, PhD2, Erika Cottrell, PhD1,2, Rachel Gold, PhD, MPH2,3

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

Introduction: There is a growing national emphasis on documenting patients’ social determinants of health (SDH) data in electronic health records (EHRs).1-2 Preliminary research suggests that EHR-documented SDH data can inform care teams about patient needs, facilitate referrals to local resources, and track referral outcomes.3-4 However, some healthcare systems may lack the capacity and / or infrastructure needed to implement systematic, EHR-embedded SDH screening, or to respond to patients’ documented SDH-related needs.5 The need for standardized SDH data collection and response is greatest in ‘safety net’ community health centers (CHCs), whose vulnerable patients have higher disease burden and worse health outcomes than the general US population.6 In 2016, OCHIN (the nation’s largest network of CHCs on a single EHR system) developed and rolled out SDH documentation tools in an EHR shared by 440 CHCs. To evaluate factors associated with the uptake of EHR-based SDH documentation, we estimated clinic-level prevalence of use of this tool, and practice characteristics associated with its adoption.

Methods: We used data from 440 primary care CHCs that were using OCHIN’s networked EHR by 6/24/2016. OCHIN, Inc., is a non-profit health center-controlled network that hosts an Epic© EHR and manages its data, including regular validation and cleaning. Data for these analyses were extracted at OCHIN. The study population includes N=345 clinics that were active OCHIN members throughout the two-year study period (6/24/2016-5/17/2018). The primary outcome of interest was a binary indicator denoting if the clinic had ever screened a patient using the EHR’s SDH tool (yes vs. no). Clinic-level independent variables of interest included clinic type and rurality. Additionally, we measured characteristics of the clinics’ patient panels, to produce clinic-level percentages of the distributions of the following characteristics for inclusion in analyses: sex, race/ethnicity, language, age, healthcare utilization, federal poverty level, and insurance. These variables were selected as potentially influential on clinics’ choice to conduct SDH screening. We used logistic regression to model SDH tool use by clinic-level characteristics and produce odds ratios and corresponding 95% confidence intervals.

Results: Among the 345 clinics that had access to the EHR-based SDH tool over the two-year study period, 41% had ever used the tool to document SDH among their patient population. Overall, CHCs serving patients in the general community had higher odds of SDH documentation than school-based health centers (p=0.009). We observed no differences in SDH documentation by clinic rurality, percentage female, or racial/ethnic distribution of the clinic’s patients. Clinics with older patients had increased odds of SDH screening compared to those with younger populations (odds ratio [OR] = 1.046, p=0.045). Clinics that had higher rates of missing data on patients’ federal poverty levels had decreased odds of SDH screening (p<0.001). Last, clinics with higher percentages of Medicaid or uninsured visits showed lower odds of SDH screening (Medicaid: OR=0.974, p=0.051; Uninsured: OR=0.948, p<0.001).

Discussion: With national calls to document patient’s SDH in EHRs, it is important to consider whether and how such screening varies across different kinds of clinics. This study reports on adoption of an EHR-based SDH documentation tool in the two years after it was rolled out in hundreds of CHCs. Under half of these CHCs used the SDH tool at all, and several clinic characteristics were associated with the odds of SDH screening tool use. These results suggest that national initiatives seeking to encourage increased EHR-based SDH documentation may need to account for the clinic characteristics that may serve as barriers or facilitators to adopting this practice change. Such context may indicate the need to provide additional, tailored support to certain clinics, to ensure increased adoption of EHR-based SDH documentation.
References

Table 1. Prevalence and adjusted odds ratios of screening for social determinants of health through an electronic health record tool over a two-year study period, overall and by clinic characteristics

<table>
<thead>
<tr>
<th>Clinic Ever Screened for SDH</th>
<th>Yes</th>
<th>No</th>
<th>Odds Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVERALL</td>
<td>N=141 (40.9%)</td>
<td>N=204 (59.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clinic characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinic Type, N (col %)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHC</td>
<td>124 (87.9%)</td>
<td>135 (66.2%)</td>
<td>0.191</td>
<td>0.009</td>
</tr>
<tr>
<td>SBHC</td>
<td>4 (2.8%)</td>
<td>29 (14.2%)</td>
<td>0.515</td>
<td>0.098</td>
</tr>
<tr>
<td>Other</td>
<td>13 (9.2%)</td>
<td>40 (19.6%)</td>
<td>1.524</td>
<td>0.425</td>
</tr>
<tr>
<td>Clinic Rurality, N (col %)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small Town/Rural</td>
<td>17 (12.1%)</td>
<td>20 (9.8%)</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Metropolitan</td>
<td>107 (75.9%)</td>
<td>159 (77.9%)</td>
<td>1.055</td>
<td>0.925</td>
</tr>
<tr>
<td>Micropolitan</td>
<td>17 (12.1%)</td>
<td>25 (12.3%)</td>
<td>1.027</td>
<td>0.216</td>
</tr>
<tr>
<td><strong>Clinic Patient-panel Demographics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female, mean %</td>
<td>56.20</td>
<td>53.95</td>
<td>1.019</td>
<td>0.111</td>
</tr>
<tr>
<td>Racial-Ethnic Groups, mean %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>47.99</td>
<td>47.03</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>16.19</td>
<td>15.17</td>
<td>1.010</td>
<td>0.190</td>
</tr>
<tr>
<td>Non-Hispanic Other</td>
<td>8.38</td>
<td>9.62</td>
<td>1.020</td>
<td>0.248</td>
</tr>
<tr>
<td>Hispanic</td>
<td>27.44</td>
<td>28.18</td>
<td>1.027</td>
<td>0.216</td>
</tr>
<tr>
<td>Language, mean %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>77.82</td>
<td>78.21</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>17.39</td>
<td>16.54</td>
<td>0.983</td>
<td>0.522</td>
</tr>
<tr>
<td>Other</td>
<td>4.79</td>
<td>5.25</td>
<td>0.980</td>
<td>0.187</td>
</tr>
<tr>
<td>Age at First Study Visit, mean</td>
<td>35.90</td>
<td>31.41</td>
<td>1.046</td>
<td>0.045</td>
</tr>
<tr>
<td>Visits per Patient, mean</td>
<td>12.73</td>
<td>13.43</td>
<td>1.005</td>
<td>0.717</td>
</tr>
<tr>
<td>Federal Poverty Level, mean %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 138%</td>
<td>19.41</td>
<td>11.84</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>&lt;= 138%</td>
<td>67.23</td>
<td>55.81</td>
<td>0.982</td>
<td>0.215</td>
</tr>
<tr>
<td>No Information</td>
<td>13.36</td>
<td>32.35</td>
<td>0.948</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Insurance Type at Last Visit, mean %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>16.64</td>
<td>14.50</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Medicaid</td>
<td>49.20</td>
<td>47.45</td>
<td>0.974</td>
<td>0.051</td>
</tr>
<tr>
<td>Medicare</td>
<td>10.59</td>
<td>7.66</td>
<td>0.980</td>
<td>0.426</td>
</tr>
<tr>
<td>Other Public</td>
<td>2.85</td>
<td>1.24</td>
<td>1.002</td>
<td>0.958</td>
</tr>
<tr>
<td>Uninsured</td>
<td>20.73</td>
<td>29.14</td>
<td>0.948</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Note: SDH = Social Determinants of Health; CHC = Community Health Center, SBHC = Social & Behavioral Health Clinic.
An Analysis of the Lexical Structure of Documented Communicative Information Surrounding the Abdominal Roentgenogram Ordered in the Neonatal Intensive Care Unit

Jacob McPadden, MS, MD\textsuperscript{1,2}, David Chartash, PhD\textsuperscript{1}, Kelson Zawack, PhD\textsuperscript{1}
\textsuperscript{1}Yale Center for Medical Informatics, Yale University School of Medicine, New Haven, CT; \textsuperscript{2}Department of Pediatrics, Yale University School of Medicine, New Haven, CT;

Introduction
Communication between clinical teams and radiologists, results in structured reports which offer a record of the remote diagnostic task of a read by the radiologists, and support the effective transmission of information.\textsuperscript{1} This communication is not a unidirectional transmission of information from the radiologist to the clinical team, rather, it consists of a study request (order) with clinical indication which prompts the subsequent report from the radiologist. To analyze the linguistic data within these reports, the communication between the clinical team and radiologist can be viewed as a conversation consisting of a question in the form of the clinical indication and an answer in the form of the radiology impression. This question and answer pair is not formulated as the typical wh-question structure, rather, consists of a clinical statement which entails the radiological answer. For example, in the case of pneumonia, a patient’s indication describing a cough results in a chest roentgenogram, which has an impression which includes the finding of a lobar consolidation concerning for infection. In this instance, it can be stated that cough entails pneumonia, by way of lobar consolidation.

In neonatal intensive care, it is commonly stated that there are a reduced set of health problems encountered. It follows that the majority of abdominal roentgenograms are obtained for a smaller subset of clinical indications. This results in the hypothesis that indication and impression pairs fall into a reduced smaller subset of categories. The presumption is that radiologists actively respond to the indication leading to related categories in response.

For a set of neonatal abdominal roentgenograms, this study aims to examine: 1. The content-completeness of communication between the clinical team and radiologist, as modeled by the language in the indication and impression. 2. The categorization of neonatal indications for a radiology study, and the correlation between these categories and impression categories.

Methods
Abdominal roentgenograms reports were obtained from all patients admitted to the Yale Neonatal Intensive Care Unit (NICU) from January 2011 to August 2018. This produced a dataset of 5149 reports. Neonates were 56.1% male with an average gestational age of 35.8(23-42). The high-level report segments defined by the consensus of the Radiological Society of North America\textsuperscript{2} were extracted. The indication gives the reason for the test and the impression gives the summary of its result.

Each extracted note segment was processed by MetaMap, extracting Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) and semantic Type Unique Identifiers (TUIs). The CUIs and TUIs in the indication and impression pairs were analyzed for internal non-exclusive consistency from indication to impression to address whether the concepts in the indication were addressed by the radiology read.


$\chi^2$ tests of association were used to test between the impression categories generated by PCA and the coded indication
categories. One test was done to identify associations between indications and impressions categories generally. A second set of tests were done between each combination of indication and impression categories.

Hierarchical clustering (using a Euclidean distance metric) was performed on the matrix of coded indication categories and impression PCA constructed categories. This was performed to analyze the separability of indications given impression categorization.

Results

Analysis of CUIs showed that only 7.7% of CUIs in the indication were referenced in the impression. By generalizing to semantic type, 44.6% of TUIs in the indication were found in the impression.

Visual inspection of PCA results initially indicated that the CUIs within impressions could be optimally grouped into between three and six classes, with the constructed dimensions adequately separating roentgenograms. \( \chi^2 \) tests demonstrated statistically significant association at four, five, and six \((\alpha<0.05)\) categories with three not having statistically significant separation. Examination of the \( \chi^2 \) test of category-category association showed significance between multiple groups. However the significance was largely driven by strong associations of the "Confirm line/device placement or positioning" categories. Analysis of dendrograms identifies optimal categorical separation off of the root node between five and six categories. These results implied six categories intrinsic to the data.

Discussion

The low overlap of CUIs between indication and impression pairs and the relative increase in the association of TUIs could be explained by the granularity of lexical concepts. This is further explicated by the mismatch between disease and radiologic finding concepts: e.g. concern for necrotizing enterocolitis vs no free air, pneumatosis, or dilated bowel loops.

The significant association between classified indications and PCA categorized impressions suggests that a conversation may be taking place. The most significant findings involved the clinical classification of devices suggesting that there was notable communication around these devices. However, this signal may be amplified by the radiologist noting the finding of a new device with or without the question being stated. It is rare in clinical experience for a radiologist not to comment on a present device and device placement confirmation is a frequent request in neonatology. Contrarily, some known information is withheld in radiology reads. A repeat roentgenogram may not repeat the unchanged findings from the previous study. This would have the effect of weakening the association.

While there may be true communication, further investigation is warranted. The difference in number between clinical categorization and PCA derived categories may indicate an sub-optimal clinical categorization. Restructuring clinical categories into more inclusive groups may improve evidence of communication. Additional approaches planned for the future analysis include textual entailment detection and topic modeling.

References


Reasons, Information, and Time: Exploration of Dental Clinician-Initiated Medical Consultations

Jayanth Kumar Medam, MS, BDS1,2, Karmen S. Williams, DrPH, MBA3,4, Jay Patel, MS, BDS1,2 Theresa Gonzalez, DDS2, Thankam P. Thyvalikath, DMD, MDS, PhD1,2,4

1Indiana University School of Informatics and Computing, Indianapolis, IN, USA; 2Indiana University School of Dentistry, Indianapolis, IN, USA; 3Indiana University Fairbanks School of Public Health, Indianapolis, IN, USA; 4Regenstrief Institute, Indianapolis, IN, USA

Introduction

Advancements in medical research have led to improved life expectancy worldwide.1 However with this improved life expectancy, approximately 26 percent of the US adult population have multiple medical conditions.2 As a result, dental clinicians are seeing an increase in patient populations with multiple medical conditions, which require both an assessment of patients’ medical information for treatment, and consideration of impacts of conditions and medications on patients’ oral health.3,4 Access to up-to-date medical information can assist in both avoiding unnecessary adverse events, such as excessive bleeding from medication, and preventing delays in oral health care. Currently, medical information is gathered in two ways—through patient-reported medical histories and medical consultations provided by the patients’ medical provider. The most common way is patient-reported medical histories; however, studies show patient-reported information has low rates of congruence between electronic dental records and electronic medical records.5 To confirm and verify patient information, dental clinicians may request medical consultations through a structured or unstructured consultation form. Although more than 90 percent of dental and medical providers use electronic dental or medical records, and there is the ability to use regional Health Information Exchange (HIE) organizations, consultations are still printed and faxed. This can cause delays in patient care and treatment. There is a pressing need to determine the challenges that delay the delivery of comprehensive oral health care by dental clinicians. Few studies have investigated the impact of medical consultations on dental treatments, importance of structured forms of medical consultations, and the characteristics of patients who require medical consultations. However, scant information exist regarding the reasons for dental clinicians to seek medical consultations despite obtaining patient-reported medical histories and to what extent providers respond to the dental clinicians’ questions. Using the Indiana University School of Dentistry structured, medical consultation forms, the objective of this study was to examine: 1) the reasons for dental clinician-initiated medical consultations, 2) the information requested and returned, and 3) days taken to complete the medical consultation process. The results of this study will inform how dentists can leverage regional HIEs to access patient medical information, such as labs and medicine.

Methods

Based on prior studies and medical consultations, two researchers drafted an initial annotation guideline. Four team members used guidelines to review the same 5 consultations with a moderate inter-reviewer agreement (Fleiss kappa-0.50). The guidelines were reviewed and finalized based on discussion and consensus. A convenience sample of faxed and returned medical consultations were selected for patient encounters between January 1, 2015 and December 31, 2017. Of these 4,586 medical consultations, 240 were reviewed for 179 unique patients. The annotation guideline characterized the reasons for the initiation of the medical consultations and type of information requested by dental clinician and provided by medical provider (Table 1). Descriptive statistics were used to analyze gender and age group. Summaries were used to determine the top reasons for initiation of medical consultation, medical information requested and received, and days of process completion.
Table 1. Sample of type of information requested by dental clinician and provided by medical provider

<table>
<thead>
<tr>
<th>Type of information</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory and diagnostic reports</td>
<td>When dental clinician requests for laboratory values and/or diagnostic reports of the patient.</td>
<td>HbA1c, WBC count, RBC count, complete blood count, viral load, etc.</td>
</tr>
<tr>
<td>Recommendations</td>
<td>When dental clinician requests recommendations or precautions or contraindications for the planned dental treatment.</td>
<td>antibiotic prophylaxis, anticoagulant medication, local anesthesia with vasoconstrictor, any contraindications, etc.</td>
</tr>
</tbody>
</table>

Results

Among the 240 medical consultations (179 unique patients), approximately 52 percent were men, followed by 47 percent women, and less than one percent listed as other. The mean age was 61 years (±15.25 standard deviation). The most common reason for dental clinician-initiated medical consultations was needing more information from medical provider (136), followed by confirmation of medical condition or treatment (103). The information requested by dental clinicians was laboratory values and diagnostic reports (219), recommendations (150), medication information (62), and the current status of medical conditions (57). However, of the 219 requests for laboratory values sent by dental clinicians, only 76 consultations returned by medical provider, followed by recommendations in 98 of 150 sent. However, medical providers returned more than requested information for medical status (85) and medication information (83), even when dental clinicians did not request. Just over half of the medical consultations were returned within 10 days with an overall mean of 15.6 days. Although, only seven were returned after 100 days, almost 15 percent of the medical consultations were sent after 30 days.

Conclusions

Although viewing a small sample size, challenges in the medical consultation process are clear. Despite studies to improve the medical consultation process, results still show gaps and delays in information medical providers share with dental clinicians. The lack of standard protocols for initiated consultation leads to duplicated efforts and delays in treatment, which affects overall health of patient. We propose that the need for medical consultations would decrease if dental clinicians had access to patients’ medical records. Access to patient medical records through regional health information exchanges and interoperability networks, such as Carequality and Common Well Health Alliance, would reduce the burden for dental and medical providers. Thus, medical consultations could be utilized to address more important questions regarding patient management. With an increase of dental patients with multiple chronic conditions, efficiently gathering medical information to deliver comprehensive oral health care is critical.

References

Automating new-user cohort construction with coarsened exact matching and
ingication embeddings

Rachel D. Melamed, PhD

1University of Chicago, Chicago, IL

Introduction

Comparative cohort studies are a major tool for assessing drug safety, but these studies require appropriate adjustment for confounding variables. Propensity score methods adjust for confounders by estimating the effect of these confounders on treatment assignment. One of the challenges of applying these methods to health care data is the high dimensionality of the health record, with tens of thousands of potential confounders. Thus, recent studies have proposed improvements to the propensity score, including large-scale regularized estimation, and automatic determination of relevant confounders. Fewer studies have focused on the selection of patient cohorts, a necessary step preceding the application of the propensity score. This step is usually designed by experts, limiting the throughput of comparative cohort studies. Including non-comparable patients can add bias and noise to effect estimates. For example, Weinstein, et al, showed that people with a history of gastrointestinal bleeding are unlikely to be prescribed paracetamol rather than ibuprofen. In order to more rapidly conduct cohort studies, we develop an automated approach to choose comparator cohorts for a treated population.

Methods

Our study uses the IBM MarketScan claims data, containing age, gender and week for coded prescriptions, diagnoses, and procedures, comprising 30,000 non-rare event types. Encoding these events results in a high-dimensional, sparse, and low-information feature vector: for example, the feature vector for a person with a diagnosis code for diabetes is not any closer to that of someone prescribed insulin than to that of a person taking Prozac. One way to create a more informative representation of medical history is with embeddings. Unlike other medical embeddings, we develop embeddings to specifically represent the medical context for new drug prescription, which we term indication embeddings. These embeddings are designed such that history events that precede the same prescription (indications) will have similar vectors to each other; correspondingly, so will drugs that occur in a similar history context.

Given any drug of interest, our method uses three steps to identify a comparator population similar to the treated population. The first step uses the drug embeddings to identify comparator drugs prescribed in the most similar medical contexts, among those drugs with the same therapeutic purpose (as annotated in UMLS). Each of these comparator drugs defines a pool of possibly comparable people. In the second step, we select from these pools of people exposed to either drug using Coarsened Exact Matching (CEM). We define coarsened bins as: gender; calendar year of prescription; age in bins of: 0-5, 6-10, 11-15, 16-25, 26-40, 41-55, 55-70; and number of unique drugs prescribed to that person. For each comparator person, that person is only kept for further analysis if that person falls into a bin containing a treated person. Third, we match patients within treated bins. Using the indication embeddings we create a low-dimensional and dense representation of patient state at the time of exposure. We match treated patients to comparator patients in each bin using the Mahalanobis distance of these vectors. Finally, to evaluate how comparable our populations are, we use the same data (demographics, medical history), to fit high-dimensional,
sparse models to predict treatment (propensity scores). We examine the propensity score overlap across the cohort definitions.

**Results**

We evaluate how comparable a pair of cohorts is using the propensity score overlap. If the propensity scores of the treated and comparator populations have high overlap then the patients have a reasonable probability of receiving either drug, that is, they are more comparable. Equivalently, propensity score models for comparable populations have a low area under the ROC curve (AUC). Drugs prescribed in similar medical contexts should have similar embeddings. First, we show that similar embeddings indeed point to more appropriate comparator drugs for a given treatment, even in the complex case of neuropsychiatric drugs (Fig. 1). Then, we assess performance of the last two steps of our method, which identify comparable populations given a drug pair. We roughly follow three published cohort studies, where the cohorts were defined by experts (Fig. 2). In each of these, we examine the overlap in the two treatment-defined populations before any matching (Fig 2, left columns); then after the CEM step (middle column); and after matching both on coarsened demographics and on history vectors (right columns). In all cases, the populations become more comparable.

**Conclusions**

We have shown that automating cohort selection has the potential to improve upon expert-defined comparator cohorts. As well, we have promising results in automatic selection of comparator drugs. This will enable cohort studies without relying on clinicians to design every study.

**References**


Feasibility of a machine learning based method to generate personalized nutrition goals for diabetes self-management

Elliot G. Mitchell, MA¹, Marissa Burgermaster, PhD², Elizabeth M. Heitkemper, PhD¹, Matthew E. Levine, BA³, Yishen Miao, MA¹, Esteban G. Tabak, PhD³, Arlene Smaldone, PhD⁵, David J. Albers, PhD⁶, Lena Mamykina, PhD¹

¹Columbia University, Department of Biomedical Informatics, New York, NY; ²The University of Texas at Austin, Dell Medical School, Austin, TX; ³California Institute of Technology, Pasadena, CA; ⁴Courant Institute of Mathematical Sciences, New York, NY; ⁵Columbia University, School of Nursing, New York, NY; ⁶University of Colorado, Departments of Pediatrics and Informatics, Aurora, CO

Introduction

Chronic diseases like diabetes are increasingly prevalent, with a high burden for individuals and the healthcare system, but this burden is not shared equally.¹ Low socioeconomic status and communities of color have more than 50% higher risk of developing obesity and diabetes compared to the overall population.² In type 2 diabetes (T2D), care involves modifying day-to-day lifestyle behaviors. Self-management is particularly challenging because of notable individual differences. In the case of nutrition and T2D, there are individual differences in glucose physiology after eating different foods, and the American Diabetes Association guidelines recommend setting individualized nutrition targets with a diabetes educator.³,⁴

Patient-generated health data (PGHD), collected through sensors and smartphone apps, offers an unprecedented view of individual physiology and behavior. Most personal informatics applications focus on visualization of PGHD, putting the burden on users to find insights that inform action.⁵,⁶ These applications require health literacy and numeracy to use, and therefore have the potential to deepen existing inequalities.⁷ When combined with machine learning (ML), PGHD can make predictions and find patterns in tasks where human cognition falls short. However, translating patterns discovered using machine learning into actionable information that could inform future choices is non-trivial, particularly when designing for populations with low literacy and low familiarity with technology.

GlucoGoalie

Here, we present the design and evaluation of GlucoGoalie, a system for individuals with T2D that recommends personalized nutrition goals to improve blood glucose (BG) control. To design GlucoGoalie we followed an iterative process that considered the data and ML methods available, expert input, a literature review of best practices in nutrition management for T2D, and early feedback from (N=3) users in the target population.

GlucoGoalie takes patient-generated data including meal logs and BG readings as input, and uses an ML-driven expert system to generate concrete and actionable suggestions for how to change eating behaviors. To deliver recommendations that are meant to inform behavior, GlucoGoalie uses goal setting, an effective behavior change strategy for diabetes self-management and mobile health.⁸,⁹ Each goal suggests changes to the macronutrient composition of meals to improve the impact on BG by specifying a direction of change and target amount. The target is presented in grams as well as “choices”, a unit used to help individuals identify portions of different foods with similar amounts of a specific macronutrient. For example, 1 carbohydrate choice is 15g, which could be 1 slice of toast or 1/3 cup of rice, while 1 protein choice is 7g of protein. See Table 1 for examples of goals generated by GlucoGoalie.

Methods

15 individuals with T2D participated in a controlled evaluation to assess whether recommendations were understandable and actionable. First, in a goal comprehension task, participants chose which of two meals followed a given goal based on meal photographs or nutritional labels. We assessed how often participants selected the appropriate meal. Second, in a meal choice task, participants were asked to assemble meals with cutout food images at a “virtual buffet.” Participants started by selecting a meal that is closest to what they would normally eat, and then were presented with two goals that asked them to vary from their baseline behavior and assemble three meals for three days in a row with
Type | Title | Goal description
--- | --- | ---
Decrease | Decrease your carbs to 2½ carb choices | For high carb lunches, decrease your carbs to be about 2½ carb choices (38g). An example of 1 carb choice is 1 slice of whole wheat toast, ⅓ cup of plantain, or ½ cup of brown rice.
Increase | Increase your protein to 3 protein choices | For low protein dinners, increase your protein to be about 3 lean protein choices (21g). An example of 1 lean protein choice is 1 ounce of lean ground beef, ½ cup of tofu, or 1 ounce of chicken breast.

Table 1: A selection of personalized nutritional goals output by GlucoGoalie

the chosen goal in mind. The procedure was repeated for breakfast, lunch, and dinner meals. We assessed how often composed meals were consistent with the direction of the chosen goal as a percentage, and tested for significance with a binomial test. In addition, we measured the variability in meeting the goal target with mean absolute error, measured in “choices”, which provides a standard unit to compare goals for different macronutrients.

Results

For the goal comprehension task, when choosing which of the two nutrition labels met a given goal participants were correct 89% (SD = 21%) of the time. When choosing between two meal images, however, participants chose the correct meal only 49% (SD = 25%) of the time. For the meal choice task, participants composed meals that were consistent with the direction of their chosen goal 67% of the time, which was statistically significant with a binomial test ($p < 0.001$). There was no difference by meal type, macronutrient, or direction of goal. While most of the chosen meals were in the correct direction, there was a high degree of variability in precisely meeting the goal target. Meals were an average of 0.83 (SD = 0.56) macronutrient “choices” away from the goal target. For example, participants were an average of 4/5 carb choices (12g), from the target.

Discussion and Conclusion

We designed, implemented, and evaluated GlucoGoalie, an ML-based expert system to generate personalized nutrition goals for individuals with T2D. Our evaluation demonstrated the preliminary feasibility of algorithmically generating nutritional recommendations from PGHD. This is important, because GlucoGoalie can help personalize self-management without requiring users to analyze their own records, or rely on their healthcare providers to identify meaningful patterns and recommend changes in behavior. However, our results indicated mixed success in participants’ ability to fully understand and follow goals, which suggests that additional work is necessary to design data-driven nutrition recommendations that are easy to incorporate into daily self-management practices.

References

Abstract

In recent years, the use of extracorporeal membrane oxygenation (ECMO) therapy has been increasing. Despite many published studies, there are no commonly accepted therapy guidelines. To address this gap in knowledge, we have developed a retrospective study of almost 900 ECMO patients to outline meaningful indications, provide clinical decision support where evidence-based recommendations can be established, and effectively counsel our patients and their families regarding the morbidity and mortality associated with this therapy.

Introduction

The number of patients with refractory cardiac or respiratory failure that receive extracorporeal membrane oxygenation (ECMO) therapy is growing. This is an expensive and life-saving technology that comes with significant morbidity and mortality. Limitations in data governance, standardization, curation, and in IT infrastructure have reduced our ability to fully utilize this information in improving healthcare delivery to ECMO patients and drive the research agenda. This project aims to streamline, standardize and integrate data from over five years of Duke experience to maximize patient sample size, clarify the proper patient selection for ECMO, better understand clinical outcomes, develop clinical guidelines for ECMO therapy, and facilitate registry submission (ELSO).

Methods

The ECMO Repository integrates data from the Duke Heart Center, the electronic health record (EPIC/Maestro Care), as well as manually curated data from REDCap on all Duke patients receiving ECMO from June 23, 2013 to October 31, 2018 (see Figure 1). The repository contains demographic information, medical history, diagnoses, medication history, medication administration, details of ECMO cannulation, pre and post ECMO lab values, vital signs, heart catheterization reports, echocardiogram reports, perfusion data, ventilation parameters, ECMO-specific risk scores, consult orders, functional assessments and transfusion requirements. The ECMO repository is built in a PostgreSQL environment and available to collaborators for analysis with appropriate approval.

Results

The current iteration of data contains 899 patients that received venoarterial (VA) or venovenous (VV) ECMO support from June 2013 through October 2018. The average age is 41.7 years of age (range neonate – 82 years of age) and 59.3% of the patients are male. The overall inpatient mortality during index hospitalization is 41.3% in the pediatric population (N=175, age < 18 YO) and 43.5% in the adult population (N=680, age >=18 YO). In the pediatric population, the most common hospital associated diagnosis was liveborn (28.9%), congenital anomalies (25.7%), related to infectious causes (15.7%), and respiratory disease (13.2%). In the adult population, the most common hospital associated diagnoses were related to pulmonary/respiratory disease (26.5%), associated with...
surgical procedures or devices (18.3%), related to infectious causes (15.1%), and complications of acute coronary syndrome (14.6%)^2.

There have been many challenges in data acquisition, cleaning, and integration of clinical data. Notably, poor clinical documentation, lack of regulatory coding schema, and clinical complexity of critically ill patients. The process of cleaning and integration is ongoing. After the complete repository is available for analysis, we plan to construct a durable pipeline to facilitate ongoing evaluation of ECMO data by automating repository updates from each clinical source every three months.

**Discussion**

This repository will allow for extensive study of ECMO therapy in Duke patients. To date, our repository is one of the most extensive and largest single-center studies of ECMO patients. We plan to use traditional (regression-based) methods and machine learning algorithms as analytic tools to build predictive models and clinical decision support systems for morbidity and mortality. This extended dataset will help us better understand which groups of patients benefit the most from ECMO therapy, the appropriate timing of ECMO therapy, and to better characterize complications of patients that have received ECMO therapy. The repository will also be available for retrospective review of ECMO patients for Duke investigators that obtain IRB and ECMO Oversight Committee approval.

![Figure 1. ECMO Project Architecture](image)

**References**


Complementing Observational Signal with Distributed Representations for Drug Side-effect Prediction

Justin Mower\textsuperscript{1}, Trevor Cohen\textsuperscript{2}, Devika Subramanian\textsuperscript{1}

1. Department of Computer Science, Rice University, Houston, Texas
2. Biomedical Informatics and Medical Education, University of Washington Medicine, Seattle, Washington

Introduction:
Contemporary approaches to identification of potential on-market adverse drug effects (ADEs) depend upon aggregation of many data sources, including spontaneous reporting systems (SRSs) and the biomedical literature. Due to inherent scale and complexity, information from these data sources require robust and scalable methods. The aim of this research is to leverage relational information extracted from the biomedical literature in direct complement to observational signal extracted from the FDA Adverse Event Reporting System (FAERS, a SRS) using representation learning and composition with downstream supervised learning for drug safety monitoring.

Methods:
Distributed semantic vector representations of UMLS concepts were derived from around 90 million concept-relationship-concept triples (semantic predications) extracted from the biomedical literature, available via SemMedDB\textsuperscript{1}. Representations were generated using a neural-probabilistic technique: Embedding of Semantic Predications (ESP)\textsuperscript{2}. Representations for drug-ADE pairs from the Observational Medical Outcomes Partnership (OMOP) and the Exploring and Understanding Adverse Drug Reactions (EUADR) reference standards\textsuperscript{3,4} were composed from these concept vectors to generate abstract “relational embeddings”, shown previously to hold promise for pharmacovigilance research\textsuperscript{5}. Additionally, disproportionality measures were extracted from Harpaz et al\textsuperscript{1} and Banda et al\textsuperscript{6} for each of the drug/side-effect pairs in the reference standards (respectively). Using each data source (FAERS and literature) individually and combined - directly as concatenated features - cross-validated average precision scores at varying training set sizes were calculated for L1-regularized logistic regression (L1LR) models. These scores were then organized into learning curves for each model/data-source combination. Additionally, an ensemble configuration was generated by training two L1LR models - one on disproportionality data alone and another on literature data alone - and calculating a weighted average. Average precision was calculated at various thresholds of literature contribution for final classification.

Results:
Across all models, incorporating information from both the disproportionality and literature data sources resulted in the highest performance, with average precision of \(-0.95\) for training sizes of \(>35\%\) of each data set, a 12-19\% relative increase over the best performing SRS-based models. Compared to literature-based models, the combination models resulted in a modest but consistent \(-2\%\) point absolute increase. For the ensemble models, the best performance of the thresholds tested was at equal contributions of the outputs of both models (that is, both literature and disproportionality models were weighted equally; a true average). At a weighted contribution of 20\% literature and 80\% disproportionality, an uplift of \(-20\%\) over utilizing disproportionality-based models alone is observed.

Discussion:
Integrating signal from the SRS and information from the literature results in improved performance over a single data source alone, demonstrating their complementation in a supervised machine learning setting. These results exceed the previous state of the art on the reference standard tested, besting the previous state of the art (even accounting for methods which consider SRS, literature, product label, and other sources of data)\textsuperscript{6}. More importantly, models presented here can be trained with varying degrees of literature contribution, allowing a user-defined level of grounding in SRS-based measures. In a conservative setting, allowing only 20\% contribution from the literature, this approach still has substantial gains over utilizing SRS-based disproportionality measures alone, the most commonly leveraged form of post-market surveillance measures.
Average cross-validation average precision scores at various training set sizes, plus or minus the standard error of the mean across multiple cross-validation runs, for disproportionality only based models (orange), literature only (green) and combined literature-disproportionality based models (blue), on the OMOP reference standard.

References:

Acknowledgements:

This work is supported by U.S. National Library of Medicine Grant (R01LM011563), Robust Inference from Observational Data with Distributed Representations of Conceptual Relations.
Framework for Clinical Decision Support and Structured Reporting in Radiology

Sujith K. Nair, B-Tech1, Mike S. Tilkin, MS2, Laura P. Coombs, PhD3
1American College of Radiology, Reston, VA; 2American College of Radiology, Reston, VA; 3American College of Radiology, Reston, VA

Abstract
Radiologists often look into standard guidelines while they dictate a case, but these guidelines are outside their dictation system, which contribute to an inefficient workflow at the point of care. ACR Assist is a clinical decision support framework designed to provide structured clinical guidance to radiologists into their workflow, also fixes the downside of narrative-based radiology reporting and ensures that every report captures information essential to patient care and valuable for analysis.

Introduction
The American College of Radiology (ACR), headquartered in Reston, Va., is a nonprofit professional society with more than 38,000 member radiologists, medical students, medical physicists and radiation oncologists. The ACR Core Purpose is to serve patients and society by empowering members to advance the practice, science and professions of radiological care.

ACR Assist is a clinical decision support framework designed to provide structured clinical guidance to radiologists in a manner that allows this content to be incorporated naturally into the radiology workflow. The ACR Assist framework includes raw clinical content, an encoding schema that allows this content to be consumed by commercial applications, and a communication framework that facilitates content delivery.

Methods
The American college of Radiology developed an open authoring for point-of-care clinical decision support tools integrated into the radiologist reporting environment referred to as the computer-assisted reporting and decision support (CARDS)1, 2 framework. The CARDS, described herein, includes: (1) a definition format for representing radiology clinical guidelines as structured, machine-readable Extensible Markup Language documents and (2) a user-friendly reference implementation3 to test the fidelity of the created definition files with the clinical guideline. The proposed definition format and reference implementation will enable content creators to develop CARDS tools that voice recognition software (VRS) vendors can use to extend the commercial tools currently in use. In making the definition format and reference implementation software freely available, we hope to empower individual radiologists, expert groups such as the ACR, and VRS vendors to develop a robust ecosystem of CAR/DS tools that can further improve the quality and efficiency of the patient care that our field provides.

A sample CARDS module creation has been explained with a figure reference (Figure 1); which shows the full cycle (1) creation of a computer readable clinical guidance (2) how a voice recognition dictation software renders it and (3) how this produce a radiology report. The framework also makes it easy to integrate AI into the workflow where any of these elements on the clinical decision support content could be filled by an AI algorithm.

Results
The framework has been adapted by prominent radiology reporting software vendors and integrated on their dictation workflow.

Figure 1. A clinical guidance converted to an XML file is rendered on a voice recognition dictation software produces radiology report.
Decreasing unnecessary variation in radiology reporting and producing guideline-concordant reports helps with the success in value-based payment models and good for patient care. ACR has been converting College’s clinical content, including the Incident Findings Committee White Papers, available for decision support tools based on the CARDS framework.

**Conclusion**

A need for a clinical decision framework for radiologists to use the clinical guidance integrated on their radiology workflow at the point of care will make them more efficient. The need for having structured radiology reports in order to help the referring physician to understand the diagnosis improves the patient care plan. Last but not the least, data sharing for registries and research will help to improve the patient care over the period of time and an open framework which facilitates the data-sharing is critical. (see: www.icmje.org/index.html).

**References**

QUICK: A FHIR Logical Model for Clinical Decision Support and Clinical Quality Measurement

Claude Nanjo, MPH, MAAS,1 Guilherme Del Fiol, MD, PhD,1 Douglas K. Martin, MD,1 Richard L. Bradshaw, PhD,1 Bryn Rhodes,2 Floyd Eisenberg, MD, MPH,3 Kensaku Kawamoto, MD, PhD, MHS1

1University of Utah Department of Biomedical Informatics, Salt Lake City, UT, 2Dynamic Content Group, Orem, UT, 3iParsimony, Washington, DC

Introduction

The Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR) data interface standard holds great potential for improving health and healthcare. FHIR is a standard for the electronic exchange of healthcare information which defines a metamodeling language, an extensible clinical information model, and a resource-oriented RESTful API.1 However, use cases such as clinical decision support (CDS), electronic clinical quality measurement (eCQM), electronic phenotyping, and cohort identification face important challenges when using FHIR directly as a target model. Specifically, CDS and eCQM systems need a consistent logical view of information (known as a logical model) that is optimized for reasoning and computation, hides unnecessary implementation details, is stable over time, and can be easily converted to and from the targeted information transport format. Here, the authors – who are leaders in the HL7 community – identify requirements for an ideal logical model for CDS and eCQM. We then propose a FHIR logical model known as the Quality Improvement and Clinical Knowledge (QUICK) model as the solution, and we evaluate QUICK and native FHIR against the identified requirements.

Methods

To develop and evaluate QUICK, we convened a panel of experts in the development and implementation of health IT standards for CDS and eCQM. The panel members serve or have served as co-chairs of the HL7 CDS Work Group (KK, GDF, BR), co-chair of the HL7 Clinical Quality Information (CQI) Work Group (FE), co-chair of the HL7 Clinical Information Modeling Initiative (CIMI) Work Group (CN), and members of the HL7 Board of Directors (KK, FE). Moreover, the panel members have been deeply engaged in current and prior efforts to develop and leverage standard logical models for CDS and eCQM, including the National Quality Forum Quality Data Model (QDM), HL7 Virtual Medical Record (vMR), the initial HL7 QUICK model, and the HL7 FHIR Quality Improvement Core (QI-Core) profiles. They have used FHIR and several logical models in the authoring of knowledge artifacts and in CDS applications deployed in production.

Drawing upon these experiences, as well as a detailed analysis of challenges associated with using FHIR directly as the logical model for CDS and CQI, the panel members developed requirements for an ideal logical model for CDS and CQI through multiple joint meetings of the HL7 CDS, CQI, and CIMI Work Groups in 2017 and 2018. QUICK was targeted for a redesign to fulfill these requirements, and it was proposed at the January 2019 HL7 Work Group Meeting as the future direction for a logical model for CDS and CQI. The approach was endorsed by the HL7 CDS, CQI, and CIMI Work Groups, and detailed data modeling based on this approach is currently underway. Moreover, to assess implementation feasibility and model fitness for purpose, the University of Utah is currently conducting a pilot study on the definition and use of declarative bi-directional FHIR-to-QUICK transformations for the automated generation of Java libraries to convert FHIR DSTU2, STU3 and R4 instances to QUICK and back for use in CQL-based, Java-based, and Drools-based CDS services. The QUICK logical model and pilot results will be submitted as a For-Comment Ballot in January 2020 by the HL7 CQI Work Group.

Results

Table 1 lists the logical model requirements developed by the HL7 CDS, CQI, and CIMI Work Groups along with the corresponding approach used by QUICK to address each requirement.
## Table 1. CDS-eCQM Logical Model Requirements and QUICK Approach to Addressing Need

<table>
<thead>
<tr>
<th>Logical Model Requirement</th>
<th>QUICK Approach to Addressing Need</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural consistency</td>
<td>Introduction of additional FHIR resource hierarchies that group resources into resource families, each sharing a common structural underpinning and reducing cross-resource variability.</td>
</tr>
<tr>
<td>Semantic consistency</td>
<td>Interface-based approach that presents optimized logical views over the FHIR model, facilitates logic operations, and results in better separation of semantic concerns.</td>
</tr>
<tr>
<td>Implementability and alignment with principles of object-oriented programming</td>
<td>Avoidance of modifying attributes that alter the semantics of the owning class. Better handling of cross-attribute dependencies and choice types. Avoidance of design-by-constraint, which often poses a challenge in object-oriented languages.</td>
</tr>
<tr>
<td>Abstraction of implementation details</td>
<td>Single way to access attribute in the model by hiding extensions, attribute slicing, and reference handling behind simple accessor methods where possible.</td>
</tr>
<tr>
<td>Model stability</td>
<td>Modular model with core module defining primitive types, complex data types, and the most stable projection of resources relevant for CDS and eCQM. Each subsequent module builds on the core module by adding additional expressivity through a design by class specialization and composition.</td>
</tr>
<tr>
<td>Intuitive to domain experts</td>
<td>Interface-based approach that supports more clinician-friendly interfaces to FHIR resources.</td>
</tr>
<tr>
<td>Expressivity</td>
<td>Modular model with core module representing the most stable projection of resources relevant for CDS and each subsequent module adding additional expressivity through design by specialization and composition.</td>
</tr>
<tr>
<td>Translation to FHIR</td>
<td>Platform independent model reversed engineered from current FHIR model to reduce transformation costs between QUICK and FHIR. Collaboration with HL7 Working Groups to improve cross resource consistency and better align FHIR to QUICK.</td>
</tr>
</tbody>
</table>

## Discussion

QUICK is designed to provide an implementable logical representation of the FHIR model for the health quality domain. The approach taken is strongly influenced by current object-oriented paradigms which are nearly ubiquitous in industry today, and it is based on substantial implementation experience in this space.

While QUICK addresses many of the implementation challenges identified over the years, it does not address all limitations equally gracefully (e.g., for the handling of FHIR “choice types”), deviates from FHIR in some areas (e.g., in the handling of modifying attributes), does not cover the entire FHIR resource space at this time, and requires restructuring transformations to and from FHIR in some cases. This inability to optimally meet all requirements relates to the inherent tension between some of the requirements in Table 1. For instance, approaches taken to enhance model semantic consistency and intuitiveness may lead to deviations from the FHIR model that complicate translations to and from FHIR. Ideally, some of the weaknesses present in the FHIR model with regard to the requirements in Table 1 will be addressed over time. In particular, the authors are engaged in a collaborative effort with other HL7 Work Groups to address structural and semantic inconsistencies in FHIR directly within the base FHIR standard. However, as the requirements for implementation models such as FHIR and logical models such as QUICK can and will differ, we anticipate a continuing need for QUICK.

## Conclusion

While FHIR aims to standardize the service interface, the need for a corresponding logical model in the quality space that lies behind the service interface has become more salient. QUICK is an attempt to present a common logical representation of the FHIR model based on implementation experience that supports both knowledge authors and CDS implementers.

## Acknowledgement

This project is supported in part by grant U24CA204800-01 from the Informatics Technology for Cancer Research (ITCR) program of the National Cancer Institute.

## References

A new representation of disease conditions and treatment pathways accurately predicts mortality and chronic diseases

Che Ngufor, PhD¹, Pedro Caraballo, M.D.¹, Thomas J. O Byrne¹, David Chen, Ph.D.¹, Nilay D. Shah, Ph.D¹,
Michael Steinbach, Ph.D², Gyorgy Simon, Ph.D²
¹Mayo Clinic, Rochester; ²University of Minnesota, Minneapolis

Introduction

The number of individuals in the U.S. suffering from multiple chronic conditions (MCC) has increased substantially and continues to rise. Patients with MCCs require significantly more health care resources, incur high cost of care, and face greater mortality, functional decline, and worse quality of life. At the basic level, the distinct MCCs act as competing risks, and co-occur either by chance, through causal relationships, or by common underlying risk factors. Consequently, interventions that address one comorbidity without reducing the severity of the others will offer limited benefit in terms of quality of life and mortality. Because slowing the progression of one comorbidity may hinder the progression of others, understanding the interactions among MCCs and treatment pathways offers the opportunity to strengthen intervention strategies. However, identifying these relationships or ascertaining the likelihood of adverse health outcomes have been hindered by the marked complexity and heterogeneity of patients with MCC.

Predictive analytics based on longitudinal data can provide crucial information for making better clinical decision about MCC patients. However, it also raises an important question about the appropriate representation of MCCs for predicting health outcomes. Numerous comorbidity measures have been developed for research using electronic health records (EHR) and administrative claims data such as the Charlson and Elixhauser Comorbidity Index (CI) and the Agency for Healthcare Research and Quality’s (AHRQ) Clinical Classification Software (CCS). Current Comparative Effectiveness Research and Evidence Based Practices (EBP) tend to focus on inefficient comorbidity representations such as simple summaries (e.g. counts, sum, frequencies, any or last observed) of individual conditions or aggregate measures of the CI and CCS within a specific observation window. However, these representations do not reflect an individual’s comorbid disease history and severity, and do not account for interactions with other risk factors or the different treatment pathways that a patient may take. This thus leaves a large gap in our knowledge about how to optimally manage individuals with complex MCCs. Different subpopulations of patients can exhibit different relationships between patient characteristics and outcomes. Thus information about the interactions and subgroups that maximally capture heterogeneity in MCCs and treatment pathways can facilitate diagnosis, enhance preventive strategies, improve quality of life, and help create smart EBP guidelines.

To address this critical knowledge gap, in this study, we introduce a novel representation of patient data called Disease Severity Hierarchy (DSH) that explores specific diseases and their known treatment pathways in a nested fashion to create subpopulations in a clinically meaningful way. As the DSH tree is traversed from the root towards the leaves, we encounter subpopulations that share increasing richer amounts of clinical details such as similar disease severity, illness trajectories, and time to event that are discriminative, and suitable for learning risk stratification models.

Methods

Study population: We used data for 15,391 adults, age 45 – 85 years, included in the Rochester Epidemiology Project (REP) database who received primary care at Mayo Clinic in 2004-2015. Subjects entered the study at their age in Jan 1 2004 and were still alive on Dec 31 2010. Patients were then followed until Dec 31 2015. Primary outcomes considered include all-cause mortality and major cardiovascular event (MCE) in 2011 - 2015. DSH construction: We focus on four common diseases: type 2 diabetes mellitus (T2DM), hypertension, hyperlipidemia and obesity and constructed their DSH for comparison with traditional representations. We included all primary care patients with or without an indication of these conditions at any time during the study period. We developed models based on features taken in 2004-2010 to predict mortality and MCE in 2011 - 2015. Specifically, for each patient, we collected demographic data, time-stamped diagnosis of the four disease conditions, medications, and laboratory results measured in 2004-2010 and use these to construct the DSH trees for each disease. For most disease conditions, associated laboratory tests exist, and medications indicated for these conditions are also known. However, most existing representation techniques ignore these relationships, thus eroding the interpretability and clinical applicability of the results. DSH is designed to account for known relationships in EHR by encoding disease severity. At each time point (hospital visit), DSH considered the full clinical context of a disease at several nested levels of details. Starting at the population level (root) we determined if a patient had a disease condition or not (e.g T2DM). If the patient is diseased, we looked for any information regarding whether the condition was treated (e.g. prescription of Metformin); next we consider the
aggressiveness of the therapy (first-line / last-line drug); and finally whether the patient was under control or not. A patient is under control, if the lab result or vital sign associated with the condition was within its predefined normal range. This nested or hierarchical representation of information helped us to assess the severity of the disease. Figure (a) shows the DSH for T2DM, where the leaves indicate if the patient is in control or not based on hemoglobin A1c. Notice that the structure of DSH is a binary tree, where each node has at most two children. Further, any branch on a right child leads to intensification of the disease severity. We used this property to quantitively represent the information embedded in DSH. Specifically, we assigned “risk” scores to the nodes, where the score of the right child doubles that of its parent. The risk score of the root node is either 1 (diseased) or 0 (disease free).

Training and validation: We developed age specific survival models, where instead of the traditional time-on-study, we used the subject’s age as the time scale to predict the risk of death and MCE at age 60, 65, 75, and 80 years based on the DSH, while adjusting for sex and race. The use of age time scale provides an expressive and flexible way to control the effect of age especially for older adults. It also provides a relatively meaningful basis on which to examine how risk varies over time.

We trained the random survival forest (RSF) model through a 5-fold cross-validation procedure.

Results

The median age of the study population in 2004 was 47.6 years, with 57.7% female and 88.5% white. 3.3% of the patients died during the follow up period (2011-2015), while 25.7% had MCE. Figure (b) presents the AUC, Sensitivity, and PPV of the RSF model trained using DSH risk scores and standard representation of the four comorbid conditions. Specifically, we considered three formats for the standard representation: a diagnosed condition persisted throughout 2004-2010, most frequent value (yes/no) of the condition, and the last observed value of the condition in 2004-2010. We also included indicators of medication use in 2004-2010. The figure clearly shows that the performance of RSF based on DSH risk scores significantly outperformed the traditional representation of comorbidities.

Conclusion

The proposed DSH risk scores effectively and accurately predict the age at which a patient maybe at risk of dying or developing MCE significantly better than traditional representation of disease conditions. DSH utilizes known relationships among various entities in EHR data to capture disease severity in a natural way and has the additional benefit of being expressive and interpretable. This novel patient representation can help support critical decision making, development of smart EBP guidelines, and enhance healthcare care and disease management by helping to identify and reduce disease burden among high-risk patients.

Acknowledgement: This study was supported by NSF grants IIS-1602394 and IIS-1602198

References

Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy

Christopher Norman, M Sc,1,2 Mariska Leeflang, PhD,2 Raphaël Porcher, PhD,3 Aurélie Névéol, PhD1

1 LIMSI, CNRS, Université Paris Saclay, 91405 Orsay, France
2 Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
3 METHODS Team, CRESS, Inserm U1153; Université Paris Descartes, 1 place du Parvis Notre-Dame, 75004, Paris, France

Introduction

The large and increasing number of new studies published each year is making literature identification in systematic reviews ever more time-consuming and costly.1 Technological assistance has been suggested as an alternative to the conventional, manual study identification to mitigate the cost, but previous literature has mainly evaluated methods in terms of workload reduction recall, i.e. the number of references missed.2 There is a need to also evaluate whether screening prioritization methods leads to the same results and conclusions as exhaustive manual screening. In this study we examined the impact one screening prioritization method based on active learning would have had on estimates of sensitivity and specificity in systematic reviews of diagnostic test accuracy.

Methods

We simulated the screening process in 48 Cochrane reviews of diagnostic test accuracy, and re-ran 400 meta-analyses based on at least 3 studies. We evaluated prioritized screening with candidate references ranked in descending order of likelihood of being relevant, trained using active learning with logistic regression on n-grams trained with stochastic gradient descent. We compared the prioritized approach with randomized order, equivalent to conventional practice for screening in systematic reviews. We examined if the screening could have been stopped before identifying all relevant studies while still producing reliable summary estimates (points estimates of the sensitivity and specificity of the diagnostic test). For all meta-analyses, we also examined the relationship between the total number of relevant studies and the reliability of the final estimates.

Results

The main meta-analysis in each systematic review could have been performed after screening an average of 30% of the candidate articles (range: 0.07% to 100%) (fig. 1). No systematic review would have required screening more than 2,308 studies, whereas manual screening would have required screening up to 43,363 studies. Despite an average 70% recall, this procedure would have changed the estimates by only 1.3% on average (fig. 1).

We found a moderately strong correlation (Pearson $r = 0.60$) between the total number of studies included in a meta-analysis and the observed uncertainty of the meta-analysis. We estimated the maximum estimation error in the reported meta-analyses to be approximately 50% point when based on 3 studies, 16% point when based on 5 studies, and 4% point when based on 10 studies.

Discussion

By monitoring the moving average of the displacement we were able to estimate the current precision of the diagnostic test accuracy estimates through the screening process. However, the meta-analyses of diagnostic test accuracy were accurate within 2% only for meta-analyses including at least 20 studies. In many cases the screening could be stopped prematurely while bounding the estimation error within prespecified limits (Fig. 1).

Future Work

Future work will evaluate the validity of these results in prospective settings. We also plan to use Bayesian methods to estimate final meta-analysis accuracy from the study data accumulated through the screening process. We will also aim to extend this approach to other study types beyond diagnostic test accuracy, such as intervention studies.
Conclusion

Screening prioritization coupled with stopping criteria in diagnostic test accuracy reviews can reliably detect when the screening process has identified a sufficient number of studies to perform the main meta-analysis with an accuracy within pre-specified tolerance limits. However, many of the systematic reviews did not identify a sufficient number of studies that the meta-analyses were accurate within a 2% limit even with conventional, exhaustive manual screening.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

Figure 1: The impact of screening prioritization and stopping criteria as the difference in meta-analysis results for the largest meta-analysis in 38 systematic reviews. Ten systematic reviews did not include any meta-analysis based on three or more studies in PubMed and were therefore excluded from the results. Effort denotes the fraction of candidate references screened. Recall denotes the fraction of identified relevant studies. Blue data points correspond to the simulated results using early stopping. Red data points correspond to results without early stopping, i.e. equivalent to current practice (which would have 100% effort and 100% recall).

References


Reduction in Severe Ordering Errors of Blood Products in Pediatrics through Formative and Summative Usability Testing

Evan W. Orenstein, MD1,2, Jeanne Boudreaux1,2, MD, Margo Rollins, MD1,2, Jennifer Jones, RN2, Christy Bryant, RN, BSN2, Dean Karavite, MS3, Naveen Muthu, MD3, Jessica Hike, MPH2, Herb Williams, BA2, Alexis Carter, MD1,2, Cassandra D. Josephson, MD, 1,2

1Emory University School of Medicine, Atlanta, GA; 2Children’s Healthcare of Atlanta, Atlanta, GA; 3Children’s Hospital of Philadelphia, Philadelphia, PA;

Introduction: Electronic order sets have the potential to improve adherence to evidence-based practices, reduce variation in care, and minimize cognitive load. However, poorly designed order sets can also change practice patterns unintentionally1 or lead to medical errors.2 In pediatric medicine, ordering blood products for transfusion is a complex task as providers must consider multiple factors, including but not limited to the transfusion itself (e.g. the appropriate blood product, volume, and rate of transfusion), patient factors (e.g. conditions requiring special preparation by the blood bank such as irradiating or washing blood products), and workflow factors (e.g. preparing blood on hold for surgery or future transfusion vs. blood to be prepared and administered immediately). During a Failure Modes and Effects Analysis (FMEA), our institution identified the blood product ordering process as high risk for serious transfusion-related errors. To mitigate this risk, we used user-centered design principles to standardize blood product ordering across a large pediatric healthcare system and evaluated order set designs using formative and summative scenario-based usability testing.

Methods: We re-designed the blood product ordering process at Children’s Healthcare of Atlanta in three phases. 1) Identification of stakeholders across clinical, laboratory, and information technology departments and assembly of a multidisciplinary guiding coalition. This group used FMEA to identify the users involved in blood product administration and nuances of blood ordering tasks by clinical department and patient context. 2) This group performed a heuristic review of existing blood bank order sets and proposed adjustments to a single common blood product ordering process. 3) Formative usability testing using a think-aloud protocol, in which providers ordered blood with the proposed order set in their usual clinical workspace using a realistic scenario appropriate to their clinical specialty. Based on observed errors and direct feedback, iterative adjustments were made to the proposed order set between participants. These processes were completed from July – December 2018.

Usability of the re-designed order set was evaluated and compared to the original version through summative testing of the effectiveness, efficiency, satisfaction, and educational value of each approach (Figure 1).

We recruited front-line ordering clinicians in their usual clinical setting from a purposeful sample of clinical specialties. Participants first answered 5 multiple choice quiz questions asking about which special processing requests would be appropriate in a complicated clinical scenario, created and validated by a pediatric hematologist (JB) and two transfusion medicine specialists (MR, CJ). Participants were then randomized in blocks of two to use either the original or re-designed order set to order blood products for 4 standardized scenarios in a non-production electronic health record (EHR) environment. Participants used the same order set version for all 4 scenarios. The scenarios were developed based on a combination of known safety events and errors observed during formative usability testing. Order set appearance and functionality based on age and weight of the patient were identical to the production EHR environment. However, clinical details of each scenario were read to the participant directly and not present within the simulated patient’s chart. For each scenario, participants were timed from the end of the scenario prompt until the orders were signed. Additionally, an observer noted any severe or moderate ordering errors, defined as follows:

Figure 1: Summative Testing Study Design

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Order Set</td>
<td>Original Order Set</td>
<td>Original Order Set</td>
<td>Original Order Set</td>
</tr>
<tr>
<td>Re-designed Order Set</td>
<td>Re-designed Order Set</td>
<td>Re-designed Order Set</td>
<td>Re-designed Order Set</td>
</tr>
<tr>
<td>Quiz (30 min)</td>
<td>Quiz (30 min)</td>
<td>Quiz (30 min)</td>
<td>Quiz (30 min)</td>
</tr>
<tr>
<td>Survey (30 min)</td>
<td>Survey (30 min)</td>
<td>Survey (30 min)</td>
<td>Survey (30 min)</td>
</tr>
</tbody>
</table>

1423
Severe error: Orders that, if followed exactly by both the blood bank and the nurse, would lead to over-transfusion (higher volume of blood or faster rate of infusion than intended reaching the patient), under-transfusion, or missed special processing request (e.g. non-irradiated blood in a scenario in which the patient has T-cell dysfunction).

Moderate error: Orders that, if followed exactly by both the blood bank and the nurse, would lead to an unnecessary special processing request (e.g. order for CMV-negative blood in a patient who does not require it) or a delay in transfusion (e.g. missing transfuse order for the nurse).

After completing all 4 scenarios using either the original version or the re-designed version, participants re-took the same 5 question quiz with the question order and answers shuffled. They completed one additional scenario with the opposite design (i.e. if they did the first 4 scenarios with the re-designed order set, scenario 5 was with the original order set). Finally, participants rated their preference between the order sets, perceived ease of use, usefulness, and efficiency of the new order set (adapted from the Technology Acceptance Model3) and overall evaluation of the new order set. All differences between the two groups were compared using two sample t-tests in R version 3.5.2. This work was deemed non-human subjects research by the IRB of Children’s Healthcare of Atlanta.

Results: Twenty-nine stakeholders were identified across clinical, laboratory, and information technology departments. After heuristic review, the standard blood product order set was restructured from groupings by similar order type into a decision tree based on the different clinical settings identified. Defaults were also adjusted based on discussion with the stakeholder group, and questions in the “prepare” and “transfuse” orders were redesigned.

Formative usability testing was completed with 27 providers from 7 specialties with 30 unique scenarios and a total of 70 tests. Key error types targeted during formative testing included transfusion volume ordering errors, errors in selection of special processing requests, and errors in communication when preparing and transposing 2 separate aliquots of blood product from the same donor unit. During formative testing, 18 design changes were made, 12 of which were considered major (e.g. changing order question logic, changing order set structure) and 6 minor (changing wording, allowing multi-select on order questions).

Fifteen front-line ordering clinicians participated in summative testing and completed 4 scenarios each for a total of 60 scenarios, with 8 participants randomized to the re-designed order set (32 scenarios) and 7 to the original order set (28 scenarios). There were no significant differences in specialty distribution or experience between the groups.

Task completion without errors and the rate of severe errors were significantly improved with the re-designed order set, while the frequency of moderate errors and time per scenario were no different (Table 1). There was a slight trend towards greater improvement in quiz grade for those using the re-designed version, but this was not significant. On a scale of -1 (completely prefer original order set) to +1 (completely prefer re-designed order set), participants preferred the re-designed order set with a mean score of 0.56 (95% CI: 0.23 – 0.89). Overall, 71% of participants agreed or strongly agreed that they were satisfied with the new system, 79% that it was easy to use, 71% that it allowed efficient task performance, and 79% that it was useful.

Discussion: In the setting of a complex, high-risk ordering task, scenario-based usability testing performed in the clinical setting demonstrated a high rate of severe ordering errors despite stakeholder heuristic review, leading to substantial changes in order set design during formative usability testing. This method of in-situ testing and immediate adjustments based on observed errors was low cost and could be performed by a small, skilled team within 1-2 weeks. Summative testing in a simulated environment demonstrated that this approach reduced (but did not eliminate) severe ordering errors in simulation with no change in ordering efficiency. While there was a modest trend towards improved learning that our study may have been underpowered to detect, reductions in severe errors appear largely independent of changes in provider knowledge about appropriate special processing requests for blood products. While we have yet to demonstrate the clinical impact of these design changes, further research after implementation in production will determine if this effort yields a true patient safety benefit.

References:
Decision-Support Needs of ECG Technicians in Inpatient Settings

Mustafa Ozkaynak, PhD1; Andrew E. Levy, MD2; Juliana Barnard, MA1; Sharon Pincus, MA1; Michael Ho, MD, PhD1,3
1University of Colorado, Aurora, CO; 2University of Colorado Hospitals, Aurora, CO; 3VA Eastern Colorado Health Care System, Aurora, CO

Abstract
Ancillary service performance affects organizational and patient outcomes. This study aims to identify decision-support needs of ECG technicians, a critical ancillary service in inpatient settings. We devised a mixed methods study including staffing data, ECG records (request time, completion time, priority) and technician semi-structured interviews/observations. Results suggest ECG completion times fall below recommended clinical times, a suboptimal result influenced by work system elements. Decision-support tools may streamline the process and significantly shorten ECG technicians' completion times.

Introduction
Timeliness is a critical component of many ancillary health services in the hospital setting.1 For instance, delays in electrocardiogram (ECG) acquisition contribute to delayed diagnosis and treatment of such critical conditions as ST-elevation myocardial infarction (STEMI).2 Decision-support tools have previously been shown to improve efficiency and workflow for providers, nurses, and other professionals3. Decision support can potentially improve ancillary services with planning and scheduling,4 but the precise decision-support needs of ancillary services (such as ECG technicians) has not been fully explored. Previous studies on decision support, highlighted the importance of congruence of technology with work system elements (e.g., person, organization, task, environment, and tools/technologies).5 The purpose of this study is to identify decision-support needs of ECG technicians, a critical ancillary service in inpatient settings. We focused on work system elements that should be considered in technology development.

Methods
We collected quantitative and qualitative data. Quantitative data included all ECG orders between 1/1/2017 and 10/31/2018 at a single academic hospital with about 700 beds. Each record included unit name, order time, completion time, and priority level (STAT, Now, Routine; maximum completion times are 15, 60 and 120 minutes respectively). We also examined staffing data. Semi-structured interviews were completed with ECG technicians (N=14, from both day and night shifts) until saturation was achieved. Interview questions focused on ECG completion processes, team communication, and work flows with probing to contextualize and explain the quantitative data. Quantitative data were analyzed to identify ECG request patterns by time of day and patient location. Qualitative data were analyzed using content analysis and frequent team debriefings to identify emerging themes.

Results
We studied 53,274 ECG orders. The 95th percentile of ECG requests each hour across the day varied between four (4-5 AM) and 12 (6-7 AM) during the weekdays. Among all non-scheduled ECGs, 39% were ordered with STAT priority. Although average ECGs completion times (about 14 minutes) could be reasonable, a significant percentage of ECGs had delayed completion times, according to current hospital standards. STAT delays (>15 min) varied slightly (23% to 33%) across unit groups while overall delays did not vary by time of day (13%-18%). Qualitative findings included (a) variability in how staff decided prioritization of multiple STAT ECG orders across a large hospital, (b) inappropriate use of STAT label for non-clinically urgent ECGs, (c) duplicative ECG orders causing technician confusion and frustration, (d) lack of alerts to notify technicians when they completed a STAT ECG outside of expected time frames, and (e) staffing issues (e.g., short staffing and demanding pace with limited breaks). Interview data supported some of the underlying work system elements highlighted in the quantitative findings. We organized these elements using systems engineering initiative for patient safety (SEIPS) model5 in Table 1.
Discussion

We describe five interrelated work system elements related to ECG acquisition and technician work. Some of these elements can be modified (e.g., priority policies or staffing levels); however, some of the elements are not feasible to modify (e.g., two-tower structure of the hospital). Informatics interventions aimed at the modifiable components should be designed in consideration of the non-modifiable elements to prevent unintended consequences.

ECG completion times can be improved through clinical decision-support systems in three ways. First, a dashboard that shows the timeliness of ECG orders and various staffing variables could be developed for administrators for better managerial planning (e.g., staffing levels). This type of decision support can assist administrative staff in decisions related to scheduling, capacity planning and policy level decisions. Second, decision-support tools embedded into electronic health records can assist clinicians in prioritizing ECG orders. This type of decision support will allow clinicians to assign an ECG order the most appropriate prioritization based on the patient’s health status and hospital policies. Third, ECG technicians, who travel between units a majority of their shifts, can be assisted by mobile decision-support tools when there are multiple, competing orders at the same time. This type of decision support systems will guide technicians which ECG order should be fulfilled first and help coordination between multiple technicians and nurses when there are competing orders.

Table 1. SEIPS-driven work system elements that affect ECG

<table>
<thead>
<tr>
<th>Work System Element</th>
<th>Findings from qualitative analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>ECG technicians’ experience</td>
</tr>
<tr>
<td>Task</td>
<td>Order time, order priority, patient location and clinical unit, repeat orders</td>
</tr>
<tr>
<td></td>
<td>by clinicians, walking long distances</td>
</tr>
<tr>
<td>Organization</td>
<td>Standards regarding acceptable completion time, technician staffing deployment</td>
</tr>
<tr>
<td></td>
<td>and chronic technician under-staffing, task sharing, pre-op protocols</td>
</tr>
<tr>
<td>Physical Environment</td>
<td>Hospital lay out, elevator availability, location of ECG machines and charging</td>
</tr>
<tr>
<td></td>
<td>requirements of machines, patient proximity to ECG machine</td>
</tr>
<tr>
<td>Tools and Technology</td>
<td>ECG machines, “Rover” technician notification system</td>
</tr>
</tbody>
</table>

Conclusion

Decision-support systems have the potential to improve ECG technician performance and ECG completion times if the development of these systems are informed by work system elements. These improvements can enhance patient safety and quality of care in the inpatient settings.

References

Objective measures for Medical Education from the Electronic Health Record

Sarada S. Panchanathan, MD,MS\textsuperscript{1}, Andrew Muth, MD\textsuperscript{2}, Pamela Garcia-Filion, PhD\textsuperscript{1}
\textsuperscript{1}University of Arizona, Phoenix, AZ; \textsuperscript{2}VA Central Iowa HCS, Des Moines, IA

**Introduction**
Hospitals nationwide have adopted electronic health records (EHRs) over the past decade. These have served as a valuable repository of information beyond their daily use by physicians for patient care. Clinical researchers use the patient data contained in the EHR, and quality officers use information about providers and their patients to create tracking dashboards for process and outcome metrics. EHR systems also provide use data for learners at all levels of medical education; this underutilized section of data provides a unique opportunity to objectively measure outcomes of both individual residents and Graduate Medical Education (GME) programs.

The Accreditation Council for Graduate Medical Education (ACGME) describes six competencies that develop from the learner stage to expert: patient care, medical knowledge, systems-based practice, practice based learning and improvement, professionalism, and communication. Recently, the ACGME recognized the need for metrics to track and improve medical education outcomes\textsuperscript{1}. To this end, medical educators are seeking objective measures for competency-based progression in lieu of solely time-based progression. Use of EHR data to guide GME training has recently been reported. In 2017, Boston Children’s Hospital reported the use of a resident dashboard with automated training metrics\textsuperscript{2}. At Phoenix Children’s Hospital, we are developing a similar resident dashboard\textsuperscript{3}. Herein we describe the use of EHR metadata from pediatric residents to create objective metrics corresponding to ACGME competencies for patient care and professionalism.

**Methods**
From an EPIC platform, 6 years of metadata were extracted for all ambulatory clinic notes authored by pediatric residents. Note types were history and physical, progress, consultation, and procedure notes. Data for unique visits and patients, as well as patient age and gender, were captured to provide the volume and measures of continuity of patient care. ICD-10 diagnosis codes for each encounter were captured to classify disease type, as an indicator of variety and volume of resident clinical experiences. Time to note completion was calculated and used as a measure for attention to professional duties. Data visualizations were created such that an individual resident can evaluate their current clinical experience, and compare themselves to their class and with graduated classes (target benchmark).

**Results**
**Patient Care:** Aggregate resident data by class on clinical experiences were visualized using a radar graph projection. To account for the finding that different diagnoses were seen in vastly different volumes, the raw data were normalized such that 100\% (red line) on the radar chart (Fig. 1) represents the target benchmark, i.e. the clinical experiences of graduated residents. Resident data are presented as percentages.\textsuperscript{†}

**Fig 1.** Patient Care: This radar chart represents the current status of clinical experiences for resident classes compared to graduated residents (100\%).
**Professionalism:** Individual resident note completion rates for the previous 3 months were calculated as the percent of notes completed within the clinic expected time period of 3 days. The average for all residents in the clinic was 71% (Fig. 2). We selected a three month time frame to capture behavior changes at each quarterly report.

![Percent notes completed within 3 days](image)

**Fig 2:** Resident data on timeliness of note completion for a 3 month period is benchmarked to the overall average of 71% for all pediatric residents (n=49). Each resident is represented by one bar.

**Discussion**
This project demonstrates the use of note metadata to develop objective measures for pediatric resident learners and educators. Individual resident data can be inserted to benchmark their clinical experiences and performance against aggregate class data. In July 2019, these data were provided to individual residents with comparison data for the first time, and will be repeated every three months to allow for data-informed goal setting. The impact of this method on resident individualized learning plans will be examined.

In addition to the immediate effects on residents, we plan to further analyze the retrospective data of graduated residents, to determine if there is a correlation between the development of expertise over time and differences in volume and variety. This method may be extended for use by other GME programs, and is a promising mechanism for comparisons within and between programs.

There were a few additional results to highlight. We observed an underrepresentation of data for diagnoses for which no testing or prescriptions are required. We expect the representativeness will improve as residents recognize that clinical experiences are measured by the appropriate use of ICD10 coding. We were also able to observe true shifts in the clinical trends, as seen by the increase in diagnoses of anemia which reflected an underlying change of the population demographics in the last 3 years (i.e. rise in refugee rates).

**Conclusion**
We developed an education report for learners and educators to obtain objective data and develop specific learning goals across the continuum. We plan to work next to combine data from disparate EHRs (Cerner, AllScripts, and Epic) within our program to capture all of our sites. We are confident that such objective measures will enable individual learners, educators, and programs to improve both undergraduate and graduate educational outcomes.

**Acknowledgments**
We would like to express our gratitude to Mark Moehling, for his tireless work on the numerous versions of the educational reports.

**References**

A deep learning-based survival model for prediction of progression in late Age-related Macular Degeneration (AMD) from color fundus photographs

Yifan Peng, PhD¹, Tiarnan Keenan, BM BCh, PhD², Qingyu Chen, PhD¹, Elvira Agrón, MA², Wai T. Wong, MD², Emily Y. Chew, MD², Zhiyong Lu, PhD¹

¹National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, United States; ²National Eye Institute (NEI), National Institutes of Health (NIH), Bethesda, MD, United States;

Introduction

Age-related macular degeneration (AMD) is a common eye condition and a leading cause of vision loss among people aged 55 and older. Based on clinical features, the disease is classified into early, intermediate, and late stages. In some people with non-late AMD, the disease progresses to the late stage slowly; in others, it advances faster and may quickly lead to loss of vision in one or both eyes. Therefore, there is a critical need to predict progression to late AMD, in order to provide useful prognostic information for patients and clinicians.

Currently, the standard clinical method of predicting progression to late AMD uses the Simplified Severity Scale, proposed following the Age-Related Eye Disease Study (AREDS)¹. A score of 0-4 is generated for the individual by grading the presence of two risk factors (large drusen and pigmentary abnormalities) in the central retina of each eye, based on color fundus photographs (CFP). Each step on the scale is mapped to a 5-year probability of progression to late AMD (in either eye). For example, if the individual has a score of 1 on the scale, the 5-year probability of progression to late AMD is 3.0%. However, using this scale requires trained specialists and is subject to human errors. Furthermore, this method is unable to incorporate other factors such as demographic/clinical information, which is recognized as important in predicting late AMD. Therefore, there is currently an unmet need for algorithms that can perform automated prediction of late AMD from CFP.

In this project, we propose a novel framework to automatically predict progression to late AMD from CFP of both eyes. Unlike previous methods², our approach takes advantage of deep learning and survival analysis in a unified framework, which, to the best of our knowledge, has not been done in the ophthalmology domain. More importantly, our model is ‘explainable’: it mimics the human grading process by detecting individual risk factors in each eye. Thus, our model closely follows the clinical decision-making process. Unlike “black-box” approaches, it allows an ophthalmologist to inspect an interpretable result.

Methods

Our framework comprises two parts (Figure 1). The first part consists of two deep learning models to classify drusen size and pigmentary abnormalities from CFPs of both eyes. Recently, the performance of deep learning has been intensively demonstrated in clinical imaging³. In our framework, we utilized DeepSeeNet, a convolutional neural network-based model that has achieved state-of-the-art performances on automated classification of AMD severity from CFP⁴. Its drusen-net detects drusen in 3 size categories (small/none, medium, and large), and its pigment-net detects the presence or absence of pigmentary abnormalities. The second part is a Cox proportional hazards model to estimate time to late AMD⁵. The purpose of the Cox model is to evaluate simultaneously the effect of several factors on the probability of a particular event. In our case, the event is progression to late AMD and the factors are (i) drusen and pigmentary abnormalities of both eyes, and (ii) demographic/clinical information of the patient.

This study used the AREDS dataset¹† collected from a 12-year multi-center, prospective cohort study of the clinical course, prognosis, and risk factors of AMD and age-related cataract. CFPs from both eyes were obtained at the

---

¹These authors contributed equally to this work.
study baseline, the 2-year follow-up visit, and annually thereafter. The gold standard annotation (severity scale score, drusen, and pigmentary abnormalities) was performed by expert human graders at the Reading Center (University of Wisconsin). Data on all participants without late AMD in either eye at study baseline/year 0 (2,246 participants) were extracted from the AREDS dataset. Of these, 434 (19%) participants progressed to late AMD by 5 years. The overall survival is thus calculated from the baseline to the year when the participant is in the late AMD stage, or censure point if still in the non-late AMD stage.

DeepSeeNet was trained on the images of the participants in the training set (59,913 images) and tested on the images of the participants in the testing set (1,396 images). All images were cropped to generate a square image field encompassing the macula and resized to $224 \times 224$ pixels. The hyperparameters were learning rate 0.0001 and batch size 32. The training was stopped after 5 epochs once the accuracy no longer increased.

The c-index was used to evaluate the performance of our framework\(^6\). It is a generalization of the AUC and ranges from 0.5 to 1, with 0.5 indicating random predictions and 1 perfect predictions.

**Results**

We first evaluated the performance of DeepSeeNet on detection of drusen and pigmentary abnormalities (Table 1). We then compared one baseline approach and two survival models. The baseline approach mapped the severity scale from the retinal specialists to the 5-year risk from AREDS report 18\(^1\). The second approach used the framework described above, and the third approach included two additional patient factors: age and smoking history. Table 2 shows that the deep learning-based survival model achieved superior performance to traditional 5-year rates with retinal specialists (0.896 vs 0.878). Our results also demonstrated that age and smoking information are helpful in the prediction (0.903 vs 0.896).

<table>
<thead>
<tr>
<th>Drusen</th>
<th>Pigmentary abnormality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.770</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.735</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.886</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.642</td>
</tr>
</tbody>
</table>

**Table 2: Performance of deep learning-based survival model.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Y1</th>
<th>Y2</th>
<th>Y3</th>
<th>Y4</th>
<th>Y5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-year rates</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.878</td>
</tr>
<tr>
<td>DL + Survival model</td>
<td>0.919</td>
<td>0.868</td>
<td>0.871</td>
<td>0.886</td>
<td>0.896</td>
</tr>
<tr>
<td>DL + Survival model + demo.</td>
<td>0.924</td>
<td>0.884</td>
<td>0.887</td>
<td>0.898</td>
<td>0.903</td>
</tr>
</tbody>
</table>

**Discussion**

To our best knowledge, this work represents the first attempt to combine deep learning with survival analyses in ophthalmology. From the deep learning perspective, we demonstrate that DeepSeeNet can achieve state-of-the-art performance for detecting drusen and pigmentary abnormalities from color fundus photographs. The subsequent survival analysis supports the idea that deep learning can enhance the accuracy of predictions of progression to late AMD. Taken together, this work demonstrates the potential of our framework to assist decision-making in both clinical practice and research environments. In future, we plan to compare our model with other traditional methods such as logistic regression and random forests.

**Acknowledgments.** This work was supported by the Intramural Research Programs of the National Institutes of Health, National Library of Medicine and the National Eye Institute. This work was supported by the National Library of Medicine of the National Institutes of Health under award number K99LM013001-01.

**References**

Get on the Same Page: Negotiating and Aligning Knowledge and Expectations between Patients and Providers through Self-Tracking Artifacts

Adrienne Pichon, MPH¹, Megan Houterloot¹, Kayla Schiffer², Emma Horan¹, Noémie Elhadad, PhD¹

¹Columbia University Department of Biomedical Informatics, New York City, NY; ²Barnard College, New York City, NY

Introduction

Patient-generated data (PGD) are emerging as the cornerstone of chronic disease management. Self-tracking tools have shown promise in supporting patients with their self-management needs, as many occur outside of doctor visits and throughout patients’ daily lives. There is limited work, however, on the use of self-monitoring and self-tracking data as means for building trust, ensuring shared decision making between patients and providers, and ultimately improving patient-centered care. There is further limited knowledge for complex chronic diseases without clear treatments guidelines and with different types of providers engaged in care. We ground our work in management of endometriosis, a multi-factorial, systemic condition impacting an estimated 6-10% of women. Endometriosis providers form multi-disciplinary care teams and patients experiment with several self-management strategies on their own. Endometriosis is an enigmatic condition with an often-debilitating impact on patients: there are no bio-markers for providers to monitor patients’ status, treatment response varies across individuals and is difficult to predict, and patients are dissatisfied with the lack of success in their care.

In prior work, we have designed and built the Phendo app (Fig 1) for endometriosis patients to self-track their symptoms, functional assessment of health status, treatments, and self-management strategies through time. The rise of FHIR technologies provides new opportunities to inform the design of self-tracking artifacts, such as interactive visualization of self-tracked data, for the goal of shared-decision making. We explore endometriosis specialists’ perceived opportunities and challenges for using these tools. In particular, we (1) identify convergent and divergent needs across medical specialties for successful encounters; (2) assess from the providers’ perspective, the ways in which they negotiate and align goals and expectations with patients, especially when both patients and providers have relevant knowledge; and (3) elicit perceived potential and limitations of self-tracking artifacts at the point of care.

Methods

Semi-structured interviews (n=10) were conducted with endometriosis specialists (2 gynecologists, 3 surgeons, 2 physiatrists, 2 pelvic physical therapists, and 1 pain specialist). Two endometriosis personas and corresponding vignettes (chronic and acute chief complaints) were described to facilitate discussion. Providers were first probed about their approach to a typical visit with endometriosis patients, perspective about shared decision making in practice, and their attitude towards using PGD at the point of care. In the second part of the interview, two prototypes for conveying self-tracked data at point of care were provided (interactive timeline and a semi-structured graphical and text summary) (Figs 1 and 2). The prototypes operate over data collected in the Phendo app. Themes were created using inductive coding of the interview transcripts. All study activities were approved by Columbia University’s IRB.

Results

A number of key themes were identified: (1) complexity of assessing patient status and goals (e.g., methods to elicit broad symptom experiences; goals of patient-specific functional activities or reduction in pain), (2) factors and strategies underpinning or hindering communication and relationship building (e.g., roles a provider can take including coach or cheerleader; patient traits including motivation and mindset), (3) approach towards treatment and selecting specific or general treatment modalities, (4) negotiation and alignment of endometriosis knowledge and expectations around treatment and outcomes between patients and providers, and (5) features of provider and information workflow to promote a successful encounter (e.g., insurance requirements and documentation; electronic health record functionalities to upload PGD). Implications of self-tracking artifacts emerged across all themes, with their potential for grounding the clinical encounter, facilitating visit structure, reviewing past treatment choices, supporting monitoring and assessment of symptoms and goals, and aligning expectations. Across specialties, providers exhibited
differences in information needs and their granularity through time (e.g., surgeons wanted the most concise summary over a large period of time and physiatrists wanted granular, daily measures).

Provider needs and perceived benefits may inform how patients engage with self-tracking. While some providers worried about continued use of self-tracking for patients, several suggested discussing "self-tracking prescriptions" as a potential encounter outcome. The tools were seen as potentially useful to build trust by validating patients’ experience and suggesting a common language for patients and providers. Our ongoing research is exploring patient needs, the interpersonal dynamics of the patient-provider relationship, and how technology may support collaboration.

Discussion

Management of chronic conditions is challenging. The enigmatic nature of endometriosis, disparate knowledge amongst different specialties, providers, and patients, and current limited guidelines for care further complicate the provision of quality care and treatment. Self-tracking data, which depict a specific patient trajectory, have the potential to help providers in their care and to act as tools for shared decision making. Our work confirms that self-tracked data can act as powerful evidence to re-align knowledge and expectations between patients and providers. The interviews can inform the design of self-tracking artifacts in the context of the patient-provider encounter.

References

Breast Cancer Screening Rates Among Family Members of Affected Relatives

Fernanda C. G. Polubriaginof, MD PhD1, Silis Y. Jiang, PhD1, Nicholas P. Tatonetti, PhD2, David K. Vawdrey, PhD1,2

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

Abstract

Delayed diagnosis of breast cancer can increase patients’ morbidity and mortality. We used a novel algorithm to identify 9,001 family members of breast cancer patients. We found that 38.5% of patients did not receive screening as recommended by guidelines and 75.2% of women over 40 with a first-degree relative with breast cancer diagnosis were not screened. Informatics methods can be used to identify patients at risk for disease development, and therefore support clinical care.

Introduction

Breast cancer is a prevalent disease, affecting 1 in 8 women in the United States. In 2019, 268,000 new cases of breast cancer are expected in the United States.1 Breast cancer is one of the leading causes of death among women. Early diagnosis and treatment is critical to reduce mortality and morbidity in this patient population. Identification of early stage tumors without metastasis leads to better outcomes. Given the high prevalence and importance of early diagnosis, the American Cancer Society2 recommends women with an average risk of breast cancer receive disease screening with mammograms starting at age 45 while the U.S. Preventive Services Task Force (USPSTF) recommends biennial mammography screening starting at age 50.3 Women considered to be at a high risk should receive early screening starting at 40 years old and, in some cases, receive additional screening with an MRI.

With the wide adoption of electronic health records (EHRs) and increased availability of clinical data in electronic format, existing clinical databases can be used to monitor overall screening rates among women and among high-risk populations. However, there has been limited work on clinician adherence to this recommendation in large academic medical centers. We undertook this study to determine the rates of breast cancer screening among women and among women with known a family history of breast cancer.

Methods

With approval from the Institutional Review Board of Columbia University Medical Center, we conducted a retrospective analysis of family members of patients diagnosed with breast visiting NewYork-Presbyterian Hospital/Columbia University Medical Center from June 2007 to March 2019. Patients with a diagnosis of breast cancer were identified using billing and procedure codes. To identify family history in electronic health records, patients’ relatives were identified using RIFTEHR (Relation Inference from the Electronic Health Record),4 a novel, validated method that used the first name, last name, phone number and ZIP code of patients’ emergency contacts to identify familial relationships. Once the relationships were identified, RIFTEHR inferred additional relationships according to family structure. The identified relationships were previously validated using both clinical and genetic data in three distinct institutions.4

Once the cohort of family members was identified, we extracted demographic information such as sex, age, race, and ethnicity from the EHR. We measured breast cancer screening rates for the overall cohort by identifying women over the age of 50 that had at least one mammogram or breast MRI. Additionally, we built a high-risk cohort by identifying women that had at least one first-degree relative (e.g., parent, child, sibling) with breast cancer diagnosis. We measured breast cancer screening for the high-risk cohort by identifying individuals that had at least one mammogram or breast MRI after the index case diagnosis date. We present descriptive statistics of the identified cohorts, along with the rate of breast cancer screening.

Results

Overall, we identified 2,368 patients with breast cancer and with familial relationships extracted by RIFTEHR. These patients had 9,001 family members in our database representing 2,237 families. This cohort of patients was composed of 11,083 (89.9%) females, with an average age of 44 years old. Among patients with breast cancer diagnosis, 97.2% where female, with an average age of 61. A total of 67 men from 54 families were diagnosed with breast cancer, demonstrating an increased risk of disease among close relatives in these families. When assessing overall screening
rates for the cohort among women over the age of 50, 61.5% received mammography screening. Among women over age 40 who had at least one first degree relative with breast cancer diagnosis, 24.8% received at least one screening test.

**Discussion**

Secondary use of EHR data can facilitate clinical research by providing not only the medical history of patients, but also their family medical history. Many EHR implementations do not store family history information in a centralized or standardized fashion. Family medical history data is often collected and stored as free-text as part of clinical notes, which makes using this information challenging. Additionally, uncertainty about the medical history of family members, as well as inaccuracies in patient recall, compound the challenge of obtaining accurate family history data.5

Our study employed a novel method for identifying families and gathering corresponding medical histories through patient-provided emergency contact information stored in the EHR. This method could be used to identify patients at risk for disease development that are eligible for additional testing. Previous research suggests similar findings for a myriad of different diseases, most notably in relation to cancer screening.6 In our study, 38.5% of patients were not screened for breast cancer, even though early diagnosis is known to decrease morbidity and mortality. Potential explanations for the low adherence to breast cancer screening may include lack of family history documentation and physicians’ choice to not pursue screening. The fact that family history is often unavailable or stored in clinical notes presents barriers to using this data in clinical encounters,7 which may ultimately result in poor screening rates. Clinicians may consciously choose to not screen certain patients; however, due to the low rate of screening adherence, we believe this explanation alone to be less probable. Future work should focus on the use of informatics solutions, such as RIFTEHR, to identify high-risk patients eligible for screening, and alert physicians of such eligibility during the clinical visit.

There are privacy issues regarding the use of the RIFTEHR method for clinical practice, and tradeoffs must be made between providing optimal care and safeguarding the privacy and confidentiality of family members’ health information.8 Notwithstanding the ethical considerations, the use of RIFTEHR for identifying familial relationships using EHR data unlocks new opportunities for secondary use of data to facilitate identification patients at high risk for disease development and support appropriate monitoring of prevention strategies such as disease screening.

**References**

5. Green RF. Summary of workgroup meeting on use of family history information in pediatric primary care and public health. In: Vol 120 Suppl 2. 2007:S87-S100. doi:10.1542/peds.2007-1010H.
Improving Delivery of Chronic Opioid Therapy (COT) in a Primary Care Internal Medicine Clinic: Quality Improvement Study

Sravya Ponnapalli\(^1\); Andrew Johnson, MS\(^1\); Stephen LaCour, MS\(^1\); Barbara Ruddy, MD\(^2\); Heather Fields, MD\(^2\); Anita Murcko, MD\(^1\); Adela Grando, PhD\(^1\); Patricia Davis\(^2\); Mark Wallace, MD\(^2\)

\(^1\)College of Health Solutions, Biomedical Informatics, Arizona State University, Scottsdale, AZ; \(^2\)Mayo Clinic, Division of Community Internal Medicine, Scottsdale, AZ

**Introduction:** According to the Centers for Disease Control and Prevention (CDC), the rate of death from prescription opioid overdose has increased by nearly fourfold (1.5 to 5.9) from 2000 to 2014.\(^1\) Misuse of prescription opioids is a significant factor as over 20% of misusers report obtaining opioid prescriptions from a physician.\(^2\) In 2016, the CDC released recommendations to mitigate risk of misuse and overdose in opioid-based treatment of chronic pain, including prescribing the lowest effective dose of morphine milligram equivalents (MME), evaluating patients every three months, checking state prescription drug monitoring program (PDMP) databases, and obtaining urine drug screening at least annually.\(^3\) This study aims to evaluate the implementation of an evidence-based chronic opioid therapy (COT) program in a primary care setting.

**Methods:** Guidelines for managing adult patients receiving chronic opioid therapy (COT) were created based on CDC recommendations, guidelines were socialized and implemented at an academic medical center primary care internal medicine clinic in September 2017. This COT program included use of an enterprise controlled substance agreements (CSA), opioid risk tool\(^4\) (ORT), standardized opioid risk assessments, quarterly office visits, annual urine drug screening (UDS), documented review of the state PDMP database at each office visit and the assignment of a primary prescriber for each patient using COT as part of personalized treatment plan. Before each visit, the Brief Pain Inventory (BPI)\(^5\) and Quality of Life (QOL)\(^6\) questionnaires were administered to assess level of pain, quality of life, and functional status using a 10-point scale. Clinician and allied health staff (AHS) satisfaction, attitudes and beliefs were assessed quarterly using a 16-item questionnaire addressing the safety, efficacy, efficiency and satisfaction regarding COT care delivery. Staff survey participants represented stakeholders throughout the clinic care delivery process, including call center representatives, medical practice secretaries, medical assistants, licensed practical nurses, team nurse leads, care coordinators, administrators, nurse practitioners, physician assistants and physicians. Physicians in training (medical residents) were excluded due to limited or no office visits with COT patients. The questionnaire employed Likert scale response categories of “Strongly Agree”, “Agree”, “Neutral”, “Disagree” and “Strongly Disagree”. For analysis, “Strongly Agree” and “Agree” were coded agreement while “Disagree” and “Strongly Disagree” were coded as disagreement. The change was tested for significance using a logistic generalized estimating equation. To evaluate changes in patient self-reported outcomes, response data from patients completing at least four rounds of questionnaires were analyzed using Friedman’s test for repeated measures. To measure the rates of adherence with the guidelines and changes in resource utilization, the institution’s Unified Data Platform (UDP), a business data warehouse which aggregates and standardizes data across systems, was queried. Queries for adherence evaluated compliance with quarterly follow-up visits (based on patient questionnaire completion), urine drug screening at least annually, completion of the (ORT), and completion of a CSA for the cohort of 217 patients enrolled in COT between Sept. 1, 2017 and Oct. 1, 2018.
Results: Fifty-one clinicians and AHS completed the baseline survey prior to implementation of chronic opioid therapy guidelines. Twelve of the sixteen questions had a significant increase in agreement across all survey intervals (p < 0.05). All five of the questions in the patient safety domain had a significant increase in respondent agreement. After guideline implementation, more staff agreed that delivery of COT was effective and patient understanding of the opioid prescription renewal process improved. There was improvement in the level of confidence in identifying patients at risk for diversion, however most respondents still felt this remained a difficult task at 12 months after guideline implementation. Regarding guideline adherence, the highest rate was seen with execution of the (CSA) (90.8%), followed by use of the ORT (80.2%) and obtaining a UDS (73.7%). However, the attainment of quarterly COT office visits, based on completion and BPI and QOL was only 37.8%. Comparing patient self-reported outcomes (BPI & QOL) before and after the COT program, there were no significant change in mean pain scores (5.87 to 5.61), quality of life (6.12 to 5.80), and functional status due to pain (5.84 to 5.96). Finally, early evaluation of utilization among a subset of 174 patients enrolled between September 1, 2017 and March 1, 2018 demonstrates that emergency department visits in the twelve months preceding initial COT visit compared to twelve months following that visit decreased from 59 to 39 (p < 0.05). There was no significant change in hospitalizations during the same observed interval.

Discussion: This observational study of a COT guideline implementation in a single-site, primary care internal medicine academic medical center practice is unique based on the concurrent assessment of staff perceptions, adherence to guidelines and patient outcomes. Clinical and support staff perceived positive changes in the safety, efficacy, efficiency and satisfaction regarding their role in the COT care delivery processes after the implementation of guidelines and the standardization of processes for local COT delivery. Although patient adherence to quarterly visits was low, post-program implementation adherence as measured by specific guideline components by providers was excellent. Guideline implementation was less impactful on the patient experience, specifically, self-reporting did not identify improvement in levels of pain, quality of life, or functional status. The decrease in the number of emergency department visits observed in the subset of patients enrolled for more than twelve months after guideline implementation is provocative. Continued monitoring, analysis, application to the full cohort and review in the context of other measures is needed. Major limitations of this study are the observational nature, lack of a control group, and observation of a single practice. This study may not be generalizable to COT delivered outside a primary care academic medical center.

Conclusion: This observational study demonstrates that implementing a standardized, evidence-based approach to COT care in an academic primary care internal medicine clinic is associated with improved staff satisfaction and an overall positive impact on COT safety, efficacy, efficiency. Early data suggest improved patient service utilization. Future work includes incorporating new clinical decision support packages, digitized patient reported outcomes and other data sources to evaluate impact on opioid-related outcomes over time.

References:

Testing the Feasibility of an Academic-State Partnership to Combat the Opioid Epidemic in Tennessee Through Predictive Analytics

Christopher Puchi, B.S., Vanderbilt University School of Medicine, Nashville TN; Ben Tyndall, Ph.D., Tennessee Department of Health (TDOH), Nashville TN; Melissa McPheeters, Ph.D., M.P.H., TDOH, Nashville, TN; Colin G. Walsh, M.D., M.A., Vanderbilt University Medical Center, Departments of Biomedical Informatics, Medicine and Psychiatry, Nashville TN

Introduction: The state of Tennessee has partnered with Vanderbilt University Medical Center (VUMC) to form the TDOH-VUMC Experience (Principal Investigators: TN, McPheeters; VUMC, Walsh). This academic-state collaboration leverages TDOH public health, analytic, and epidemiology expertise with clinical applied predictive modeling expertise at VUMC to combat the opioid epidemic killing an average of three to four Tennesseans every day. It includes a Memorandum of Understanding and Data Use Agreement for de-identified data sharing developed over the course of one year to support analytic work as well as learning experiences and regular meetings across sites.

This study presents a preliminary predictive model to predict fatal opioid-related overdose to test this collaboration and its feasibility for data sharing and scientific collaboration. This model aggregates de-identified data including: > 110 million prescriptions in the Controlled Substance Monitoring Database (CSMD), Tennessee’s Prescription Drug Monitoring Program (PDMP); vital statistics including Medical Examiner certified cause of death; demographic data; pharmacy/provider data. We also bridge external data sets like the Area Deprivation Index Data (ADI) to improve potential for analytics to identify those at highest risk of overdose before harm occurs.

Methods: Retrospective case-control study. Data were collected from various TDOH data sources from January 2012 to December 2017: Prescription-level data from the CSMD; Hospital discharge data from state databases; State vital statistics and mortality data including Medical Examiner-certified causes of death. A major challenge at this scale remains disambiguation of patient identifiers in the CSMD. Like other PDMP’s, pharmacists type information manually for each prescription. Ongoing TDOH efforts have shown notably improved accurate patient identification, but the iterative nature of this work resulted in a subset of charts that were difficult to disambiguate. We selected only those with clear identifiers that matched across all study datasets and opted not to include those records not yet fully disambiguated for this feasibility study.

Fatal opioid overdose cases were defined as decedents in the year 2017 with drug overdose as an underlying cause of death on the death certificate (ICD-10 codes X40–X44, X60–X64, X85, and Y10–Y14) with any opioid as a contributing cause of death (ICD-10 codes T40.0–T40.4, T40.6). Control candidates were defined as those with any opioid prescription in 2017 who did not have a recorded opioid overdose death. Demographic information was obtained from death certificate data for cases and the TDOH Hospital Discharge Data system for controls. Address of residence data were obtained from the CSMD. Exclusion criteria were: 1) no demographic data; 2) no Tennessee address; 3) incomplete prescription data; and 4) multiple records of death occurring in different years.

The validation strategy split study data into random partitions of 60% for training, 20% for validation, and 20% for a holdout testing set. An optimal ratio of 1 case to 8 controls was determined empirically by evaluating model performance on validation data. Predictors were extracted from individuals’ 2017 CSMD opioid prescription records and from demographic data (Figure 1). Median ADI and county fatal overdose rate were determined using each of an individuals’ Tennessee zip codes listed in the CSMD. Age was calculated as a continuous variable reflecting age at day of death for cases and age on December 31st, 2017, for controls. Buprenorphine prescription data was obtained from the CSMD. Predictions were calibrated using logistic calibration. Relative variable importance was measured through node impurity. The experimental setup included R version 3.5.1 and packages caret and rms.

Results: We identified 1,330 opioid overdose decedents in 2017 (1,268 of these individuals had Tennessee as their state of residence on the death certificate). Of these 1,330 decedents, 858 filled opioid prescriptions in 2017. 160 individuals were excluded because their records were not fully disambiguated. 1 individual was excluded due to multiple death records. 11 individuals were excluded because their prescription records were incomplete, and 21 were excluded because they did not have a Tennessee zip code listed in the CSMD. We identified 1,977,281 individuals with opioid prescriptions in 2017 who did not have a recorded opioid overdose death. 852,066 of these individuals were excluded due to lack of demographic data, and 255,086 individuals were excluded because they did not have a

* Because of the preliminary nature of these analyses, prevalences presented might not match exactly the published Prescription Drug Overdose Report previously published by the state of TN DOH.
Tennessee address listed in the CSMD. The dataset used for model training, validation, and testing consisted of 665 cases and 5,320 randomly sampled controls from the pool of 870,129 candidates. 47 cases (7.1%) and 229 controls (4.3%) filled buprenorphine prescriptions in 2017.

A random forest model with 50 trees outperformed logistic regression on the validation set. The Area Under the Receiver Operating Characteristic Curve (AUC) was 0.83 for random forest and 0.81 for logistic regression – though formal hypothesis testing was not performed for this pilot study. Brier scores were 0.07 and 0.08 for random forest and logistic regression, respectively. A random forest model was tested on the holdout set. The AUC was 0.81 with a Brier score of 0.08. Relative variable importance is summarized in Figure 1. Logistic calibration on random forest predictions yielded an AUC of 0.837 and a Brier score of 0.08.

Conclusions: We tested the feasibility of the data sharing and modeling pipeline for the TDOH-VUMC Experience and were able to validate promising initial predictive models of fatal opioid-related overdose risk. In this example, PDMPs offer promising data repositories to identify risk factors for and risk of opioid-related overdose. Using machine learning on PDMP data may support automation of screening for risk of fatal opioid overdose. Future directions include development using larger cohorts from the PDMP, refinement of feature selection and inclusion of larger numbers of relevant predictors (e.g., buprenorphine therapy), validation of these findings on out-of-sample Tennessee data, inclusion of other external data sets such as the Community Health Service Indicator Data6, and testing other algorithmic assumptions empirically, e.g., other parametric/non-parametric algorithms, and deep learning. Simultaneously, we are understanding current clinical processes for intended implementation of this methodology with state collaborators. If effective, these findings might apply to other states using a PDMP, and we believe this collaborative model generalizes to other states willing to reach a data use agreement, to align around an educational initiative, and to consider ethical and legal issues as they arise. As the TDOH-VUMC partnership evolves, identification of objective criteria could help drive evaluation and continuous improvement efforts. The collaborative methodology described has a wide range of potential applications, from screening high-risk individuals for earlier intervention, to providing decision-support for opioid prescribers. The partnership approach also has the potential to provide rapid collaborative opportunities to address future public health challenges.

References
Extracting Drug Exposure Epochs and Drug Response Outcomes from Electronic Health Records

Andrea H. Ramirez, MD, Yaping Shi, MS, Elliot Fielstein, PhD, Jonathan Schildcrout, PhD, Henry H. Ong, PhD, Josh C. Denny, MD, Josh F. Peterson, MD
Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

Estimating drug exposures and defining their temporal relationship to drug response outcomes is challenged by incomplete prescription data and lack of reconciliation between drug exposure records. We devised a method to visualize all sources of medication data, including concept matched drug mentions, e-prescribing records, and manually curated drug lists, maintaining temporal relationships, and annotating with drug response phenotypes. The method has utility for empirically creating rules to define drug exposure epochs for each patient.

Introduction

Ascertaining medication exposure and defining temporal relationship to drug response outcomes from EHR data is challenging due to incomplete data capture from prescribers outside the system and lack of integrated pharmacy fill information. Often a simple binary classification is used for drug exposure where a patient is defined as exposure given any drug mention prior to the phenotype of interest. This simplistic approach leads to high rates of misclassification as the complex time varying nature of drug exposure and dose relative to outcome events is lost.

As part of our work with the Implementing Genomics in Practice (IGNITE) network to study the pharmacogenomic relationship of adverse events from long-term proton pump inhibitor (PPI) therapy we developed a model to represent complex drug exposure data. Defining PPI exposure is especially challenging as the drug is available over the counter as well as by prescription. Research has shown that PPI use is associated with Clostridium difficile colitis (C. diff), pneumonia, acute kidney injury (AKI), and chronic kidney disease (CKD). To characterize the additional risk of these events to patients who metabolize the drug poorly due to their CYP2C19 genotype it is important to accurately classify which subjects were most likely to be on PPI, at what dose, and at what time.

This work describes a longitudinal method for describing drug exposure relative to outcome events that was developed and built into a reusable R function including graphical representation and production of files for further analysis. Demonstrated here, using as much information as possible allows accurate exposure ascertainment, less misclassification, and improved analyses.

Methods

Data from the VUMC Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT) project including CYP2C19 genotyping for approximately 15,000 patients and their clinical data was used to study how metabolizer status impacts adverse events associated with long-term PPI usage. Computable phenotypes were implemented for C. diff, pneumonia, AKI, and CKD including extracting date(s) of events. PPI drug name, date of mention, and dose were extracted using MedEx, a natural language processing (NLP) system that recognizes medication information from any free-text record. In this example, we used problem lists as the source of our drug information.

To handle missing dose information in drug data, we used last observation carried forward (LOCF) to impute missing dose with the following schema: 1) Use LOCF to impute dose for subsequent mentions with no dose; 2) If there are mentions with no dose before the first ever observed dose, a default dose of 20mg was assigned. Drug exposure epochs are defined as periods of time during which a patient was exposed to a drug at a given dose. We
defined the start of an epoch as the date of a new PPI mention and the end of an epoch as either: 1) The date of a subsequent mention with a different dose OR 2) No mention of PPI in subsequent problem lists. At the end of a patient’s record, the last epoch is extended by 3 months from the last mention with dose or imputed dose.

Patient ID and demographics, phenotype event date, and epoch start/end dates with dose serve as inputs for our R function that will generate a graph and other files such as table of case counts with PPI exposure. A simulated graph is shown in Figure 1 with annotations to explain salient features.

Results

Two example plots are shown below in Figure 2. In the bottom row, all three events are labeled as on drug. The top row has a pneumonia event labeled as on drug, while the C. diff event is labeled as off drug. This difference is notable as the C. diff event would have been labeled as on drug when using a traditional binary definition of exposure.

![Figure 2. Example drug exposure epoch graphs.](image)

Table 1. Comparison of drug exposed case counts

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Epoch derived exposure</th>
<th>Traditional exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Diff</td>
<td>79</td>
<td>152</td>
</tr>
<tr>
<td>CKD3</td>
<td>393</td>
<td>915</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>204</td>
<td>363</td>
</tr>
<tr>
<td>AKI</td>
<td>861</td>
<td>1545</td>
</tr>
</tbody>
</table>

Discussion

This abstract describes a method to visualize information regarding drug or other exposures to outcomes for study of association such as drug metabolizer status. While the absolute numbers of exposed subjects may decrease, the certainty of the exposure is higher. This approach also facilitates sensitivity analyses to better understand the threshold around any found signal. Also, by creating simple imputation rules, more dose information may be used, as many drug-outcome associations are sensitive to dose thresholds but tracking of dose is inconsistent across the EHR. Many medications have a known starting dose, and reasonable minimum and maximums expected. While the sample demonstrated uses data extracted by the Medex NLP program, other sources of exposure for medication or other variables may be used as shown in the input format example. Other data such as insurance fill data and pharmacy fill records exist, though unfortunately this information is not routinely shared to the research community nor has proper permissions in place to reassociate with clinical notes and outcomes. Additionally, drugs such as PPI which are easily taken over the counter are better captured in provider encounters than other sources. This method is an important step forward in improving the specificity of exposed case and control assignments, and will avoid the misclassification and loss of power from simpler methods of assigning drug exposure.

Acknowledgements

This work has been funded by U01 HG007253 through NHGRI. The dataset(s) used for the analyses described were supported by institutional funding: 1S10RR025141-01 instrumentation award, and UL1TR000445 from NCATS/NIH.

References

Multi-technology Meta-Analysis Identifies Bacterial Signatures Delineating Clinical Response to Immune Checkpoint-Inhibitor Therapy in Melanoma Cohorts

Jayamary Divya Ravichandar, Cheryl-Emiliane Chow, Connor Skennerton, Dhwani Haria, Erica Rutherford, Helena Kiefel, Joanna Dreux, Kareem Graham, Kathryn Iverson, Lynn Yamamoto, Michelle Lin, Michi Willecoxon, Nicole Narayan, Paul Loriaux, Roberta Hannibal, Sabina Lau, Shoko Iwai, Sunit Jain, Thomas Weinmaier, Yonggan Wu, Yuliya Katlinskaya, Toshikko Takeuchi, Karim Dabbagh, Todd DeSantis
Second Genome Inc., South San Francisco, California.

Introduction
The gut microbiota has emerged as an important player in cancer pathology and a growing body of evidence supports its influence on clinical response to cancer therapy, especially in the context of immune checkpoint inhibitors (ICI). While several studies present insight into the landscape of bacteria differentially abundant in responders and non-responders to ICI, they are limited by biases arising from variations in the cohorts, manner in which data is procured, processing workflows, and analysis pipelines. Further, bacterial species signatures are rarely concordant across independent datasets, which can limit therapeutic development.

In this context, microbiome-based discovery of an enhancer of clinical response to ICI rests on identification of bacterial functions conferring a host benefit. Mining the genomes of differentially-abundant bacterial strains or differentially-abundant proteins in metagenomes enables identification of such functions and presents a promising approach to discover novel bacterial-derived therapeutic molecules. Herein, existing published datasets which investigated links between the gut microbiome and response to ICI were re-analyzed and integrated via a multi-technology meta-analysis (MTMA) to identify strains and proteins separating responder and non-responder populations in melanoma that are concordant across cohorts. Bacterial polypeptides derived from these response-associated signatures were screened for immune cell binding using phage display to identify polypeptides, followed by chemical synthesis of binders and evaluation in vitro for therapeutic potential to enhance clinical response to ICI.

Methods
We performed a literature survey to identify studies investigating the effect of the baseline gut microbiome on clinical response to ICI (anti-PD1) in melanoma patients. Studies were included in the analysis if both metadata and sequence data was available in public repositories. Three studies that had analyzed the microbiome associated with stool collected from discrete melanoma cohorts prior to ICI therapy were selected. Metadata was evaluated for accuracy and manually re-annotated using a controlled vocabulary, allowing for direct comparison between studies. Samples were sequenced on one or more technologies (16S NGS and shotgun metagenomics: a total of five datasets) and FASTQ files were downloaded from public repositories, quality filtered and matched to our in-house strain database (StainSelect: secondgenome.com/platform/data-analysis-tools/strainselect) to obtain strain-level annotations. For shotgun metagenomic sequencing datasets, we de novo assembled genes and predicted proteins. Based on the distribution and sparsity of each dataset, normalization and statistical tests were applied to identify strains and proteins whose abundance differed significantly between responders and non-responders. A random-effects model was used to integrate strain- and protein-level differential abundance results from all datasets. Significance is defined at adjusted p-value < 0.05. Proteins were predicted from the genomes of response-associated strains and pooled together with those significantly enriched in responder metagenomes. Proteins longer than 50 aa were broken down into smaller fragments in silico to obtain polypeptide fragments 8-50 aa long to enable efficient incorporation into phage libraries, where these polypeptides were expressed as a fusion protein on phage virions. These libraries were panned against multiple immune cell types to identify peptides that bound specifically to the target immune cells.

Results
Leveraging our multi-faceted bioinformatics platform, which enables appropriate technology-specific quality filtering and statistical testing, we identified differentially abundant bacteria at the strain-level. A range of 50 - 900 strains were identified in analysis of each dataset. No overlap in significantly differentially abundant strains was observed across cohorts or even between datasets from the same cohort sequenced on different technologies. From the de novo assembly, a range of 2-3M proteins were identified in each dataset. Of those, a few thousand proteins were significantly enriched in responders in each dataset. We performed a cross-cohort multi-technology meta-analysis (MTMA) to identify strains and proteins that were concordantly enriched in responders across datasets. Figure 1
demonstrates the power of the MTMA approach to identify strains with consistent effects across multiple datasets while eliminating strains with discordant or otherwise inconsistent effects. The strain signatures identified by MTMA were not previously reported in the analysis of these datasets in their publications1-3. The MTMA approach identified 37 strains and 1,000+ proteins as concordantly enriched in responders across cohorts. From these response-associated strains and proteins, we identified 50,000+ proteins and polypeptide fragments.

**Figure 1.** MTMA identifies concordant microbiome signatures delineating response to ICI. Significantly enriched strains (blue dots) in the responders or non-responder by individual-dataset analysis (grey dots) or MTMA (purple dot). Solid lines, supported by MTMA; dashed lines, supported by individual dataset analysis. Thick lines, significant findings; thin lines, non-significant findings. Red lines, enriched in non-responders; green lines, enriched in responders.

We hypothesized that an immunomodulatory-effector molecule that has the potential to modulate ICI-response would need to bind to immune cells associated with anti-tumor immunity. We established a phage display screening platform and screened the polypeptides against four immune cell types. Phage-panning selected for about 1% of the polypeptides as immune cell binders.

**Discussion**

Although individual datasets identify certain strains to be beneficial or detrimental with high confidence, it could be unduly biased by the cohort and platform4. The cross-cohort multi-technology meta-analysis presented here demonstrates the power of a robust bioinformatics platform to leverage publicly available datasets and systematically integrate microbial shifts (strains and proteins) to identify polypeptides with therapeutic potential. This MTMA approach on strains not only identifies concordance across cohorts, but also enables integration of datasets generated by different technologies at fine-grained bacterial identities (strain-level), resolving a key challenge with microbiome meta-analysis. Strains identified by MTMA, in addition to providing a starting point for identification of peptides with favorable therapeutic effects, could also serve as predictive biomarkers for response of melanoma patients to ICI. To our knowledge, this is the first time a phage display platform has been used to systematically screen peptides derived from bacterial signatures for immune cell binders. The platform presented here represents a novel approach to microbiome-based drug discovery, with a focus on bacterial-derived effector molecules rather than species as therapeutic targets. With the identification of concordant strain biomarkers for response to ICI and polypeptides with therapeutic potential to enhance response to ICI, this works presents a paradigm where administration of therapeutic polypeptides could be determined by the differential abundance of these bacterial signatures.

**References**

Learning Disease Phenotypes with Semi-Supervision

Victor A. Rodriguez, MPhil, Adler Perotte, MD, MA
Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

In this work we propose the Semi-Supervised Mixed Membership Model (SS3M), a probabilistic graphical model which utilizes a relatively small number of disease labels to learn multiple disease-specific phenotypes from clinical data. SS3M does not require that all patients have labels for all diseases. Thus, SS3M reduces the amount of labeled data needed to train effectively, addressing a major limitation of fully supervised phenotyping methods. Furthermore, SS3M establishes a one-to-one relationship between disease labels and disease phenotypes. Thus, unlike with fully unsupervised phenotyping methods, SS3M may be guided toward learning specific phenotypes desired by the user.

Methods

SS3M is a model for multisource bag-of-words data accompanied by a set of instance specific labels. It is closely related to the popular topic model Latent Dirichlet Allocation (LDA)\(^1\), as well as its multisource\(^2\) and supervised extensions\(^3\).

Model Description: Let \( D, S, \) and \( P \) be the number of patients, clinical data sources and phenotypes. Each patient, \( d \in \{1, \ldots, D\} \) is associated with a set of clinical observations \( \{w_{sdn}\}_{n=1}^{S} \), as well as a set of partially observed binary labels \( \{A_{dp}\}_{p=1}^{P} \). Each observation, \( w_{sdn} \), is accompanied by a latent assignment \( z_{sdn} \) which specifies which phenotype-token distribution, \( \phi_{sp} \), the observation was drawn from. Assignments for patient \( d \) are drawn from a patient-phenotype distribution, \( \theta_{d} \). A patient’s label set directly impacts her patient-phenotype distribution, and thereby all her assignments. When \( A_{dp} = 1 \) the phenotype \( p \) is “on” for patient \( d \); the model uses \( B_{p} \) to calculate \( \theta_{d} \). When \( A_{dp} = 0 \), the phenotype is “off”, and the model uses \( B^{*} \) instead. Since \( B_{p} \) is generally much larger than \( B^{*} \), setting \( A_{dp} = 1 \) encourages \( \theta_{d} \) to allocate more probability mass for phenotype \( p \).

Label sets are partially observed. If a specific label, \( A_{dp} \), is observed, then its value is held fixed. If \( A_{dp} \) is unobserved, then the model learns its value during inference. In this latter case, \( A_{dp} \) is modeled as a binary variable drawn from a distribution parameterized by \( C_{p} \) – a latent variable controlling the likelihood of phenotype \( p \) being “on” within the patient population. Note that the partial observation of labels is what makes SS3M a semi-supervised model.

Inference: We implement a collapsed Gibbs sampler to obtain posterior estimates of our model’s latent variables. We integrate the \( \{\theta_{d}\}_{d=1}^{D}, \{\phi_{sp}\}_{s=1, p=1}^{S,P} \) and \( \{C_{p}\}_{p=1}^{P} \) out of the joint distribution. The complete conditional distributions for the \( \{z_{sdn}\}_{n=1}^{N_{sd}} \) and \( \{A_{dp}\}_{d=1, p=1}^{D,P} \) in the collapsed joint are sampled from directly. We use Hamiltonian Monte Carlo to sample from the complete conditionals of the \( B_{p} \) and \( B^{*} \). Setting the path length \( L = 100 \) and step size \( \epsilon = 10^{-3} \) yielded stable trajectories with high acceptance rates in preliminary experiments.

Dataset: We train our models using clinical notes, labs, and medications for 46,520 patients from the Medical Information Mart for Intensive Care version III (MIMIC-III)\(^4\). Patients’ label sets are derived from the 50 most common ICD9 diagnosis codes. The corpora and labels were split into training and test sets containing data for 80% and 20% of patients respectively.
Experiments: We train SS3M and a closely related unsupervised model, the Multi-Channel Mixed Membership Model (MC3M), setting $P = 70$. For SS3M this results in 50 labeled and 20 unlabeled phenotypes. Preliminary studies indicated inclusion of some unlabeled phenotypes resulted in better performance for SS3M.

Qualitative Evaluation: A clinical expert assessed the quality of SS3M phenotypes relative to MC3M phenotypes by evaluating their coherence, granularity, and label quality. Coherence and granularity were assessed as described in Pivovarov et al. To assess label quality, the expert was asked to rate how well a SS3M phenotype’s true label appeared to match the learned phenotype using a 5-point scale (1-no match, 5-perfect match).

Quantitative Evaluation: For SS3M, we estimate each test patient’s posterior likelihood for each disease label. We evaluate AUROC and AUPRC against logistic regression (LR) and naive Bayes (NB) models trained with raw tokens or patient-phenotype distributions, $\{B_d\}_{d=1}^D$, learned with either MC3M or SS3M trained without labels (MC3M with a structured prior – MC3M-SP). Because SS3M is highly sensitive to the relative values of the $\{B_p\}_{p=1}^P$ and $B^*$, we train our model in two modes: (1) sample these variables freely and (2) hold them fixed.

Results

<table>
<thead>
<tr>
<th>Coherence</th>
<th>Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>SS3M 20% 8% 14% 36% 22% 24% 32% 44%</td>
<td>MC3M 18% 28% 32% 20% 2% 42% 42% 16%</td>
</tr>
<tr>
<td>True label match phen.?</td>
<td>True label match expert’s?</td>
</tr>
<tr>
<td>1 2 3 4 5</td>
<td>Y N</td>
</tr>
<tr>
<td>SS3M 36% 16% 12% 22% 14% 26% 74%</td>
<td></td>
</tr>
</tbody>
</table>

(b) Qualitative Results Summary

<table>
<thead>
<tr>
<th>SS3M</th>
<th>MC3M-SP</th>
<th>MC3M</th>
<th>Raw Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC</td>
<td>0.719</td>
<td>0.720</td>
<td>0.865</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.299</td>
<td>0.249</td>
<td>0.427</td>
</tr>
</tbody>
</table>

(e) Quantitative Results Summary

Discussion

SS3M is a model for semi-supervised learning of disease phenotypes from clinical data. We exposed SS3M to MIMIC-III clinical data and derived phenotype labels from common ICD9 diagnosis codes. In quantitative evaluations, SS3M was generally outperformed by baselines in predicting disease labels on held out patient data. Meanwhile, in qualitative evaluations conducted by a clinical expert, SS3M phenotypes were judged to be more coherent, and more granular than phenotypes learned with a closely related unsupervised model. These results suggest that despite its weak performance on label prediction, SS3M does indeed appear capable of learning interpretable, disease-specific phenotypes from clinical data. Ongoing work is focused on improving SS3M’s predictive performance and acquiring higher precision labels than the ICD9 codes used here.

References

Developing Cancer Informatics Tools to Support the Implementation and operation of a Cancer Care Navigation Program across Eight VA Facilities

Traudi Rose RN, MBA1, Jianji Yang PhD1, Judy McConnachie MPH2

1VA Portland Healthcare System, Portland, Oregon, USA
2Providence Health & Services, Portland, Oregon, USA

Introduction
The concept of patient navigators in cancer care started in 1990 with a program developed by Dr. Harold Freeman at Harlem Hospital [1]. Since the 90s, cancer navigator programs have continued to grow throughout the United States [1]. Since 2012, Patient Navigation Process is a standard of care for all programs accredited by the Commission on Cancer [2]. In October 2013, the VA Integrated Service Network (VISN) 20 Cancer Care Navigation Team (CCNT) was initiated as a pilot program across eight VA facilities. CCNT is a multidisciplinary team approach that focuses on identifying and eliminating barriers to care and coordinating care to assure that patients have timely access to care and needed psychosocial support. VISN20 covers a broad region including Alaska, Idaho, Oregon and Washington. There is a CCNT at each of the eight facilities, with all of the teams forming a VISN20 CCNT network to support Veterans traveling across the system for cancer care. Due to the unique technological barriers which include separate Computerized Patient Record System (CPRS) at each facility and need for system-wide patient tracking and communication, it was determined that informatics would be a critical component in supporting navigation. To assure close collaboration with the informatics team, the VISN20 CCNT RN Program Manager was embedded with the informatics team during CCNT program implementation and participated in every step of the cancer informatics tool development lifecycle.

Methods
Collaboration with stakeholders: An important part of the CCNT program implementation was development of cancer informatics tools. An integrated team approach enabled the embedded CCNT Program Manager to establish close relationships with the informatics team and serve as the liaison between the developers and end users. This model ensured the team had ready access to the Program Manager during the agile development process. The Program Manager also regularly updated the team on CCNT program activities and real-time feedbacks, promoting ownership and pride amongst the informatics team. The relationship established during this development process highly benefited the ongoing support and maintenance of the applications, due to team engagement and ownership.

Tool development: The Program Manager worked closely with the developers through all stages of requirement gathering, prototyping, development, usability testing, data validation, and end user training. Tools are developed using an agile approach and iterative cycles that depend on timely user feedback, facilitated by the Program Manager.

Implementation of tools: Once the clinical tools were available, training and further information was provided to the CCNTs, feedback and questions were sent through the Program Manager who serves as the interface between the clinical staff and the informatics team.

Monitoring and evaluation: CCNT’s use of tools is monitored to determine effectiveness, identify further needs, and identify under-utilized elements for removal. Information from tools is also used for the overall CCNT program evaluation and presentations to leadership and key stakeholders.

Results
Various tools have been developed and/or customized to support the operation of the CCNT program. Table 1 shows challenges to navigation and tools to meet these needs. With these tools, CCNT has now supported over 11,586 Veterans with cancer and their caregivers/loved ones (Nov. 2014- June. 2019).

Program Implementation
Care Management Tool (CMT): Web-based task management tool to support CCNT team-based care across VISN for complex patient needs. Figure 1 depicts the heavy usage of the CMT by CCNT in six facilities actively using the tool.
Standardized Documentation: CPRS consult, note titles, and templates developed and standardized for CCNT use across VISN20. Health Factors which are customized data elements to capture clinical information are embedded in templates. New Cancer Findings Report: Application pulls data daily for all facilities, to display new malignancies from pathology reports for CCNT review. VISN20 Resource Directory: Provides service-level, clinic-level and individual staff contact information across the VISN to support CCNT communication needs.

Program Evaluation
Activity Hub: A real-time data dashboard of CCNT activity. The dashboard pulls information for all CCNTs from standardized note titles, consults, clinics and CMT. Program Outcomes Data: Program Manager designed,
implemented and analyzed studies with informaticians to evaluate program objectives and impact, as well as report to VISN leadership.

Discussion
Navigating Veterans across multiple facilities is a challenge due to system-wide technological barriers. Our model of having the CCNT Program Manager embedded within a clinical informatics team has allowed for integration, development, and timely enhancements of clinical informatics tools, which has supported the reduction in barriers to care at both a system-wide and individual level. One of the challenges we faced is that the documentation tools depended on CACs to load them at each facility which caused some delays. We also found that there are varying degrees of adoption and use of tools across sites. Further data and assessment are needed to better understand the impact the navigation program and supporting cancer informatics tools have on timeliness and access to cancer care.

Table 1. Identified barriers/challenges to system-wide navigation, tools developed/customized to support these needs and specific navigation activities supported.

<table>
<thead>
<tr>
<th>Identified Barriers/Challenges</th>
<th>Tools Customized/Developed</th>
<th>Activities Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracking complex needs, across full cancer continuum</td>
<td>Care Management Tool</td>
<td>Continuity of care, workload efficiency, patient safety, patient tracking, case management, facility-team communication and coverage</td>
</tr>
<tr>
<td>No system to identify new cancer diagnoses</td>
<td>New Cancer Finding Report</td>
<td>Identification of new cancer diagnoses from all VISN20 facilities, timely notification of results to Veteran, timely development of care plans, enhanced safety and quality outcomes</td>
</tr>
<tr>
<td>Timely communication with accurate points of contact across and within facilities</td>
<td>VISN20 Resource Directory</td>
<td>Communication and coordination of care across VISN20, identification of points of contact at all facilities</td>
</tr>
<tr>
<td>Lack of clear visible ownership of coordination of care activities</td>
<td>CPRS Standardized Documentation</td>
<td>Documentation of coordination of care activities led by CCNT, standardized CCNT assessment—barriers to care identification</td>
</tr>
</tbody>
</table>

Figure 1. Care Management Tool Usage. Number of Veterans actively tracked (average 1,657 per quarter) vs. total CCNT patients and number of actions (including create, review and complete tasks, average 9,957 per quarter) taken by VISN20 CCNT since the third quarter of 2017.

References
A Visual Consideration of Sepsis Deaths and Non-sepsis Deaths with a Finely Graduated Measure of Patient Acuity: Should We Just Focus on Who is Sick?

Michael J. Rothman, PhD a; Kathy Belk, BA a; Robert Fogerty, MD b

Affiliations: aPeraHealth, Inc., Charlotte, NC; bYale New Haven Health System, New Haven, CT

Introduction
Two widely publicized numbers have had profound influence on initiatives to improve the quality of care in hospitals. The first is the statement that 30% of US hospital deaths are associated with sepsis [1]. The second states that each hour of delay after the inception of septic shock in initiating effective antimicrobial therapy increases mortality rates by 7.6% [2]. The first establishes sepsis as a major problem. The second is a call to action: early intervention can reduce mortality. We posit that these numbers have been misunderstood. Even though sepsis is a problem, a more effective approach would be to simply focus on sick patients. Although patients who arrive at the hospital with sepsis are sicker than those who do not [3], those who die of sepsis are those who arrive in poor physiological condition, just as non-sepsis patients who die are those who arrive in poor physiological condition. The Rothman Index (RI) is a finely graduated general measure of patient acuity based on empirically derived univariate relationships between 1-year post-discharge mortality and each of 26 clinical measurements and has been validated in many studies [4]. It is applicable across all diseases and settings.

Methods
Clinical data covering 27,453 inpatient visits over a 4-month period (4/16 – 8/15/2018) were extracted from the electronic medical records of three sites which are part of the Yale-New Haven Hospital System. These data included the initial RI score for each visit, along with the ICD-10 codes, and the patient’s discharge status. Univariate comparisons between septic and non-septic patients were analyzed using Chi-square tests. Comparative cumulative percentage histograms were generated for sepsis present on admission (POA) and non-sepsis POA patients discharged alive, and for sepsis POA and non-sepsis POA patients who died. Mean and standard error were computed by bucketed RI (0-20, 20-40, 40-60, 60-80, 80-100) for mortality in sepsis POA patients and non-sepsis POA patients. One and two variable logistic regression (LR) models were created, using initial RI, and then adding a sepsis POA flag. IRB approval was granted.

Results
Of 27,453 inpatient visits, 974 did not have an RI score and were excluded from the analysis. Of the remaining visits, 1,894 were designated septic POA using ICD-10 codes. The mortality rate for septic POA patients was 8.6% compared to 1.4% for non-septic POA patients (p<.001). For non-septic POA patients who were discharged alive, 20% arrived with a RI below 65 versus 55% of septic POA patients. For those who expired, the percent of non-septic POA patients with an initial RI below 65 was 86% versus 89% for septic POA patients. LR AUCs for one and two variable models were 0.907 and 0.909, respectively. The graph of mean mortality +/- 2SE vs. mean bucketed RI was similar for septic and non-septic patients (Fig. 1).

Discussion
To calibrate the reader, an RI value of 100 represents “unimpaired”. A value of 65 represents the acuity of a patient discharged to a skilled nursing facility, a value of 40 would warrant consideration for ICU transfer. Mortality rates for patients whose scores fall below 20 are about 25%, and who fall below 0, 50%. Viewing histograms in Fig. 1 we see a marked contrast in the initial RI score between septic and non-septic patients who are discharged alive. However, when considering patients who expired in the hospital that difference disappears. Non-septic patients who expired arrived at the hospital looking just like septic patients who expired. This observation is reinforced by the LR results, where initial RI alone is an excellent predictor of mortality and adding an indicator of sepsis POA provides no improvement. The graphs of mortality vs. initial RI further confirm that initial acuity, how sick the patient is, determines likelihood of death. Sepsis is largely irrelevant. In other words, to address hospital mortality rates, focus should be on patients who arrive in poor physiological condition. A recent paper by Rhee [5] goes further. After reviewing charts of septic patients who expired, they found that more aggressive treatment, specifically the measures in the sepsis bundle, would only have prevented 4% of deaths. Reducing mortality is difficult, but rather than a focus
on sepsis, a more general approach, addressing those patients who arrive at the hospital in poor physiological condition, is more likely to produce significant results. [6]

REFERENCES


Figure 1 – Top left: cumulative comparative histogram shows the distribution of initial RI scores for non-sepsis vs. sepsis patients discharged alive, top right: for patients who expired. Graph on the bottom shows mean mortality and 95% CI versus mean bucketed RI for patients with sepsis at admission and without sepsis at admission.
More Than Copy-Paste: Content-Importing in Clinical Documentation

Adam Rule, PhD\textsuperscript{1,2}, Michael F. Chiang, MD\textsuperscript{1,2}, Michelle R. Hribar, PhD\textsuperscript{1,2}

\textsuperscript{1}Department of Ophthalmology, \textsuperscript{2}Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR

Introduction

Content importing technologies such as copy-paste, text-importing phrases, note templates, and data-links provide an efficient way to document clinical encounters but can unnecessarily lengthen notes and propagate errors\textsuperscript{1}. A crucial step in weighing the risks and benefits of content importing technologies is understanding their current use and impact on clinical documentation. While there is growing evidence about the widespread use and safety risks of using of copy-paste\textsuperscript{2-5}, other methods of content importing and their impact on documentation have been largely overlooked with only a few studies reporting the percent of note text that has been imported, copied, or manually entered\textsuperscript{4,5}. The aim of this study is to demonstrate a method of using Electronic Health Record (EHR) activity logs to examine use of content importing technologies in greater detail.

Methods

This study was approved by the Institutional Review Board at Oregon Health & Science University (OHSU), which granted a waiver of informed consent for analysis of EHR logs. OHSU is a large academic medical center in Portland, Oregon. This study was conducted at the Casey Eye Institute, OHSU’s ophthalmology department, which includes over 50 faculty providers who perform more than 130,000 outpatient examinations each year. In 2006, OHSU transitioned to an institution-wide EHR (EpicCare; Epic Systems, Verona WI) and documentation at Casey Eye Institute is performed in an ophthalmology specific module (Kaleidoscope).

Our institution’s EHR tracks use of several content-importing technologies, including \textit{text-importing phrases} which import preset text into the note and \textit{data-links} which import structured data from other parts of the patient’s record. For example, typing the “.SIGN” phrase into a note and hitting enter might import a pre-set signature line. Alternatively, invoking “.VITALS” might import a patient’s most recent vital signs, using a data-link embedded in the phrase’s definition (e.g., @LASTVITALS@) to retrieve and place vital sign values into the note. Text-importing phrases in our institution’s EHR may also have embedded \textit{manual entry flags} (i.e., ***) that make it easy for providers to tab between specific locations in the note so they can manually enter text, such as exam findings or their plan.

We collected data on all text-importing phrases used while documenting office visits to 48 providers at Casey Eye Institute between Jan 1, 2017 and Dec 31, 2018. We manually coded each unique phrase as generating either 1) a full-note template, 2) a signature or attestation line, or 3) serving some other purpose. We counted the number of data-links and manual entry flags embedded in each phrase. To inspect text imported by data-links in more detail, for each of the 50 most frequently used data-links we randomly selected 100 notes generated using a text-importing phrase with that data-link embedded within it and manually counted the number of words imported by that data-link. All analyses were performed in Python (v. 3.7).

Results

Of the 200,695 office visits in 2017-18 to the 48 study providers, 96% had at least one text-importing phrase used during the visit. There were 5,810 unique phrases used 647,524 times across all visits. Full-note templates and signature lines were a small portion of unique phrases (11% and 8% respectively), but together made up more than half of all phrases usages during the study period (27% and 28% respectively).

The median office visit used three text-importing phrases, one full-note template, one signature line, and one importing text for some other purpose. These other phrases served many purposes including inserting a short phrase (e.g., “right eye”), inserting a long phrase (e.g., standard findings for a review of systems), inserting a note subsection template (e.g., a template for results of a specific exam), generating patient education materials, or generating a letter template. Phrases generating full-note templates were used at 82% of office visits. Whereas the median other phrase had 4 words of static text and the median signature phrase was 22 words, the median full-note had 81 words. Moreover, whereas the median other and signature phrases had no data-links or manual entry flags, the median full-note template had 18 data links and 4 manual entry flags. See Figure 1 for an example of part of one full-note template generated by invoking a text-importing phrase.
Figure 1: Portion of a full-note template with inline data-links (e.g., @NAME@) and manual entry flags (i.e., ***).

There were 301 unique data-links embedded in the 5,810 text-importing phrases and the amount of text imported by data-links varied significantly both within and between links. For example, whereas the links for the patient’s @AGE@ and @SEX@ only ever imported one word, the link for their @MEDS@ imported 100±77 words, sometimes importing no text and other times importing over 350 words.

Discussion

This study has three key findings: 1) content importing phrases are used during the vast majority of office visits, 2) phrases that generate full-note templates are used at the majority of office visits, and 3) full-note templates include many data-links to automatically import text. These findings highlight the frequent use of text-importing phrases to generate documentation and reliance on full note templates to automatically populate large portions of the note, rather than build it up piece by piece with several smaller phrases. These findings have implications for the design of EHRs as well as the management and use of content-importing technologies. They show the impact template design can have on note usability, with a few phrases generating templates being used for documentation at the majority of office visits. Moreover, they reveal documentation practices that resemble programming in that providers manipulate a few abstract symbols representing large blocks of text rather than writing that text out manually.

This research demonstrates using EHR activity logs to investigate use of content-importing technologies for clinical documentation. Future work can provide further guidance for the design and use of content-importing technologies by investigating the impact of specific phrases and data-links on documentation length, quality, and usability. Clinicians and administrators should think carefully about the design and use of note templates, particularly ones that include long strings of boilerplate text or large tables automatically populated with information from the patient record such as medication lists. Such templates should be designed with an eye towards what is clinically relevant and strike a balance between freedom for clinicians to generate their own templates while providing guidance and governance to help templates embody best practices.

References

Complexities of Finite State Machines to Streamline Inpatient Sepsis Management
Sameh Saleh, MD, Samuel McDonald, MD, Mujeeb Basit, MD, MMSc
UT Southwestern Medical Center, Dallas TX, US

Introduction: Sepsis is the dysregulated response to an infection that can lead to life-threatening multi-organ dysfunction. It has an annual incidence of 1.7 million cases claiming 270,000 lives and costs US hospitals nearly $25 billion per year (1,2). Despite previous research to gauge sepsis risk, the heterogeneity of sepsis makes it a difficult process to predict and preempt. Previous sepsis prediction models have had limited real-world use due to a lack of a flexible design implementation framework. These models either do not integrate into the electronic health record (EHR) or do not provide real-time recommendations that harmonize with a physician’s workflow and decrease cognitive load (3,4).

A finite state machine (FSM) is a mathematical model that defines a set of states with logical transitions linking them. FSMs are one design solution that have been often used in computing and have been successfully applied to other medical computing domains (5,6). FSMs can serve as a discretized, simple framework to define risk using a prediction model where the risk strata from the prediction model serve as the FSM transitions. The FSM can then be easily adjusted for different prediction models and use cases while directly comparing model performance and providing integrated, pertinent recommendations, both in real-time. We extend on our previous work of using an FSM to model sepsis risk in the emergency department (7) by tailoring the FSM with the same prediction model applied to the inpatient setting. We then evaluate the prediction model’s performance of sepsis to the previously used Systemic Inflammatory Response Syndrome (SIRS) criteria.

Methods: We conducted an observational study of adult (≥18 years old) inpatient (including intensive care unit) hospitalizations (>48 hours) from October 2018 to February 2019 at a single tertiary care hospital in north Texas. The finite state machine stratifies risk into Low, Medium, and High risk before and after administration of broad-spectrum antibiotics. We implement a previously validated logistic regression prediction model developed by the electronic health record (EHR) vendor Epic into our EHR to produce model output in 15-minute intervals for every patient (Figure 1). The model outputs serve as the transitions between states and are compared with standard SIRS criteria at the same interval. We first analyzed the descriptive statistics of the High state in our population and examined the amount of time patients stayed in the High state. We then focused on the High to the High + Antibiotics transition by analyzing the first transition to a High state in a given admission. The primary outcome was the positive predictive value (PPV) of that first High transition for treatment of sepsis, which was defined by the use of broad-spectrum antibiotics that are started within 24 hours of the High transition and are continued for more than 72 hours. We also calculated the negative predictive value by evaluating the Low to Low + Antibiotics transition. We initially verified the veracity of this treatment of sepsis definition in a subset of 40 patients per transition evaluated and plan to automate the process for all other admissions.

![Finite State Machine Diagram](image1)

**Figure 1.** Our finite state machine for sepsis risk for an admitted patient using a previously validated logistic regression prediction model developed by the electronic health record (EHR) Epic to define transitions. As a comparison, we use the same FSM framework, but use SIRS to define transitions. The focus is on the grayed transitions for the current study.
Preliminary Results: For 7,946 unique admissions, there were 4.85 million transitions, of which only 0.16% were from a lesser state to a High state. There were 5,016 admissions that only had Low to Low transitions throughout the entire hospitalization, while 1,421 admissions had at least one transition to a High state. Figure 2 shows a histogram of the continuous time spent in the High state every time a new High state was encountered; a vast majority exited the High state quickly within less than 4 hours. Initial chart review of 40 high-risk patients showed a 97.5% PPV for sepsis treatment when using the Epic model compared to 67.5% when using ≥ 3 of 4 SIRS criteria and 90% when ≥ 2 of 4. Negative predictive value (NPV) for a subset of 40 low-risk patients was 92.5% with the Epic model versus 40% with < 2 of 4 SIRS criteria.

Conclusions: We are able to easily extend a workable, live FSM from the emergency department to the inpatient setting. Preliminary results comparing the EHR prediction model versus SIRS criteria show clear overall gain in discrimination with improvement of both PPV for high-risk patients and NPV for low-risk patients. We expect these results will be sustained as they are validated on the full dataset.

Only a small percentage of patients enter the high-risk state and only a minority of those persist there for more than 4 hours. Establishing trend signatures that pre-date a transition to a high-risk state may help identify the risk of sepsis earlier. Determining which patients persist in the high-risk state versus those who are only transiently there will be a key future insight into what drives a septic process as well as appropriate sepsis treatment in the inpatient setting after they have received antibiotics.

Overall, FSMs allow for robust interpretability and modeling flexibility as well as real-time analysis and clinical decision support of sepsis risk in the inpatient setting. In doing so, they can decrease cognitive load and streamline workflow for physicians.

References:
A Digital Biomarker for Detection of Benign Childhood Epilepsy with Centrotemporal Spikes (BECTS)

Aarti Sathyanarayana, PhD1,2, Rima El Atrache, MD3, Michele Jackson, BA3, Ivan S. Fernandez, MD3, Kenneth D. Mandl, MD, MPH1,2, Tobias Loddenkemper, MD1,3, William J. Bosl, PhD1,2,4

1 Computational Health Informatics Program, Boston Children’s Hospital, Boston, USA, 2 Department of Pediatrics, Harvard Medical School, Boston, USA, 3 Department of Neurology, Boston Children’s Hospital, Boston, USA, 4 University of San Francisco, San Francisco, USA

Introduction

Benign childhood epilepsy with centro-temporal spikes (BECTS), also known as Rolandic Epilepsy, is one of the most common forms of focal childhood epilepsy. Despite its prevalence, it is often misdiagnosed or missed entirely. While most research has focused on early warning signs of seizure onset, a more difficult challenge is to monitor brain electrodynamics in order to detect functional characteristics - biomarkers - that indicate the brain has entered a dynamical state in which seizures are likely to occur. This has the potential to predict clinical outcomes. Our methodology uses nonlinear analysis to identify a digital biomarker for BECTS at the population level. This biomarker appears to have properties necessary for a clinically effective screening method on an “awake” EEG.

Seizures associated with BECTS occur most frequently during sleep. They present with unilateral tonic and clonic features, and at times generalized tonic clonic seizures. Because these seizures occur during sleep, BECTS often goes undiagnosed. Moreover, BECTS usually remits before the age of 18. Due to the high rates of remission during adolescence as well as the long-term effects of anti-epileptic medication, children diagnosed with BECTS often go untreated. Despite the name, this benign epilepsy has been repeatedly associated with developmental delays. The decision of whether to treat BECTS with anti-epileptic drugs or not remains unclear.

Methods

Our methods leverage nonlinear EEG signal features extracted from multiscale entropy and Recurrence Quantitative Analysis (RQA) on the EEGs of 22 BECTS patients, and 22 controls, to find reliable predictors of BECTS diagnosis.

Study Population: All patients were seen at a tertiary epilepsy center in the Division of Epilepsy and Clinical Neurophysiology, Department of Neurology, Boston Children’s Hospital (BCH). Twenty two BECTS patients between 4 and 16 years old were retrospectively chosen from the BCH Epilepsy Center database after BECTS was confirmed via clinical history and the EEG findings of an experienced neurologist. Twenty two age-matched controls with normal EEG (i.e. no diagnosis of epilepsy) were selected from the same clinical database.

Data Collection: Thirty second EEG samples were selected for both the BECTS cases and the controls. On the BECTS cases, the 30-second samples were selected such that they contained no spikes or evidence of epileptiform activity from awake EEG data. Each sample was collected on 19 channels located according to the standard 10-20 system. These segments were classified as ‘awake” by an experienced EEG technologist and reviewed by a resident neurophysiologist. All EEG samples that were chosen from awake subjects, both BECTS cases and controls, appeared normal to the neurophysiologist on visual inspection. Some of the BECTS case sleep segments contained nocturnal epileptiform spikes. No other filtering was performed on the EEG signals.

Signal Processing: Two key sets of dynamical values were computed from the EEG signals: Multiscale Entropy and Recurrence Quantitative Analysis. Multiscale entropy is a nonlinear method that has been used for analyzing biological signals and distinguishing healthy from pathological states. Another approach to computing nonlinear time series properties is recurrence quantitative analysis (RQA). RQA is an empirical approach to analyzing time series data and is in principle capable of characterizing all of the essential dynamics of a complex system including real-world, noisy, high dimensional data. In principle, RQA is capable of detecting significant state changes in a dynamical system, which suggests that it may be appropriate for detecting BECTS. All nonlinear features used in this study were computed on multiple frequency bands using publicly available methods and tools.
Results
Statistically significant group differences between BECTS patients and controls were found for several nonlinear features, particularly determinism, \( p < 10^{-4} \) for all frequency bands. Determinism is a measure of dynamical stability and non-randomness, which has been hypothesized to be related to seizure risk. The below heatmap illustrates the area under the curve (AUC) of a multiscale curve for determinism, on the awake brain of a BECTS patient versus a control. The irritative zone indicates the region of the brain that BECTS patients have identifiable epileptiform activity at night. The BECTS patients have no identifiable epileptiform activities when awake.

**Figure 1:** Brain Dynamic Effects on a 10-20 System Modeled by Determinism

Discussion
Our results suggest that although BECTS patients are most likely to seize at night, that significant dynamical differences can be detected as digital biomarkers even when awake. Our findings suggest that our nonlinear signal analysis approach may be able to enable routine clinical EEG measurements during awake interictal periods, to screen for BECTS.

References

1454
Optimizing Workflows in the Patient Portal to Incorporate Patient-Reported Outcomes in Clinical Practice

Courtney Segal, BA¹, Elizabeth Austin, MPH¹, Denise Chang, MD¹, Eliza Sutton¹, MD, Danielle Lavallee, PharmD, PhD¹
¹University of Washington Medicine, Seattle, WA

Introduction

Patient-reported outcomes (PROs), or data reported directly by patients about their health and quality of life, are used in a variety of clinical settings to provide insight on patients’ health experiences and their perspectives on health outcomes. Health systems are increasingly looking to leverage digital tools to efficiently use PROs and drive healthcare delivery, yet paper-based approaches remain the mainstay of collecting data from patients. Engaging patients outside the care setting using patient portals can introduce workflow efficiencies, improve response reliability, and enhance data quality through long-term follow-up between clinical encounters. However, little evidence exists to guide care processes and ensure appropriate clinical follow-up of PROs when using patient-facing technology, like portals. This study reports on learnings from three pilots to collect PROs via the patient portal for Medicare Annual Wellness Visits (AWV), depression screening in primary care, and symptom tracking for depression management.

Methods

The three prospective pilots took place 2017-18 at 20 primary care clinics within Seattle, WA and its surrounding suburbs (Table 1). A multidisciplinary team compromising clinicians, clinic support staff, IT programmers, population health managers, and researchers established standards and technical configurations in the EHR (Epic) to collect and report the HRA and PHQ using the portal (MyChart). The project goal focused on creating tools that facilitated more integrated data collection and efficiencies in the clinical workflow processes. We utilized quality improvement methodology (Plan Do Study Act, or PDSA) to continuously evaluate current state and rapid cycle implementation of PRO workflows. The HRA and PHQ pilots were conducted sequentially, which allowed for iterative translation of learnings through the PDSA cycle. Qualitative interviews (n=15) with patients provided user feedback on completing the assessments in the portal, including usability of the questionnaire in the portal and impact on care experience. Additionally, the team sought feedback on usability and workflows from clinic teams (n=4) via in-person meetings.

Table 1. Portal PRO Workflow and Population Summary

<table>
<thead>
<tr>
<th>Case Summary</th>
<th>A) Medicare AWV</th>
<th>B) Depression Screening</th>
<th>C) Depression Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Goals</td>
<td>Annual assessment of preventive care needs</td>
<td>Annual depression screening in primary care encounters</td>
<td>Symptom tracking for depression management</td>
</tr>
<tr>
<td>Questionnaires Administered</td>
<td>Health Risk Assessment (HRA): 45 items covering psychosocial and safety risk</td>
<td>Patient Health Questionnaire (PHQ): 2- or 9-items, valid for screening and diagnosing depression symptoms</td>
<td>PHQ-9: reliably measures response to treatment and remission among patients with depression</td>
</tr>
<tr>
<td>Population</td>
<td>Patients with Medicare FFS or Medicare Advantage scheduled for a Medicare AWV with their primary care provider</td>
<td>Patients scheduled for a routine encounter with their primary care provider, without a documented depression screening within the past 12 months</td>
<td>Patients with depression for whom the PHQ is used as a primary measure in their clinical management</td>
</tr>
<tr>
<td>Workflow Approach</td>
<td>Centralized manual attachment of HRA to upcoming appointments</td>
<td>Automated attachment of PHQ-2 or -9 to upcoming encounters based on health maintenance logic</td>
<td>Multi-team manual outreach to individual patients</td>
</tr>
<tr>
<td>Implementation Metrics</td>
<td>Messages opened; questionnaires submitted; patterns of data missingness</td>
<td>Questionnaires submitted; Distribution of scores, including high risk responses</td>
<td>Questionnaires submitted; Distribution of scores, including high risk responses</td>
</tr>
<tr>
<td>Pilot Dates</td>
<td>June-October 2018</td>
<td>August-December 2018</td>
<td>August-December 2018</td>
</tr>
</tbody>
</table>
Results

Table 2 presents a summary of questionnaires sent and submitted across the three workflow types.

Table 2. Electronic PRO Deployed and Submitted by Clinic and Workflow Type

<table>
<thead>
<tr>
<th>Questionnaire Metrics</th>
<th>A) Medicare AWV</th>
<th>B) Depression Screening</th>
<th>C) Depression Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sent (n)</td>
<td>260</td>
<td>1478</td>
<td>232</td>
</tr>
<tr>
<td>Submitted (%)</td>
<td>44.6%</td>
<td>22.3%</td>
<td>37.1%</td>
</tr>
</tbody>
</table>

Over the course of the HRA pilot, 260 patients (ages 49-103) received the HRA at three clinic sites. Sixty-nine percent of patients opened the E-mail alerting them to the questionnaire, and 45% of patients submitted the HRA prior to their AWV, with no patterns of data missingness. One clinic site participated in the depression screening and management pilots, and providers from 7 additional clinics used the PHQ for depression management with their patients. During the pilot, 232 patients received the PHQ via the depression management workflow, which had a higher response rate (37%) compared to the depression screening workflow (22.3%), which was sent to 1,478 patients.

Feedback from patients who were sent the HRA highlighted that it was easy to use and helped them better prepare for their appointment, yet recommended more targeted reminders for questionnaire completion. For the PHQ, patients highlighted that completing the PHQ streamlined their appointment allowing for more time to discuss responses with their provider. Due to the sensitive nature of answering questions about depression symptoms, patients raised the importance of communicating the purpose of depression screening in primary care, as well as personalized correspondence from a known provider for depression management.

Feedback from clinical teams using the HRA emphasized that portal workflow significantly increased clinic efficiency and allowed providers to proactively identify care needs, yet recommended more intuitive tools to track submissions. Feedback from providers participating in the depression screening and management pilots emphasized the need for clearly defined roles and clinical pathways to respond to PHQ submissions, particularly for high risk responses. Validation of the portal screening and management workflows was also important to clinical teams to ensure that PHQ data was appropriately reviewed and stored for population-based management and contractual reporting purposes.

Discussion

Our pilots of HRA and PHQ collection in the portal demonstrated feasibility for obtaining and integrating patient data prior to and between clinical encounters that supports workflow efficiency and improved patient-provider engagement. The differences in PRO response rates across the three workflows (45%, 22%, and 37% respectively) highlights the variation introduced by local clinical workflows (e.g. staff available to deploy and track questionnaires) and patient populations (e.g. preference and comfort to use the portal). Through PDSA we were able to identify changes in the technical workflow and staff training that led to increased response rates (e.g., adding E-mail reminders). Response rates may further increase with targeted and personalized reminders, as well as tools to track questionnaire completion for providers and clinic staff that check-in and room patients for their appointments. Patients in the depression care setting also emphasized the importance of communicating the purpose of screening and use of PHQ results. Additional barriers to adoption of the workflows included fitting the PRO data flow into existing and familiar clinical tools, as well as the appropriate data storage location that meet the various clinical follow-up and administrative reporting needs. Providers also desired clinical decision support to triage submissions, particularly for high risk responses. During the pilot monitoring, additional metrics were identified that would be important to track in future implementations to understand population reach, workflow efficiency, and variation in clinical pathways.

Conclusion

While promising, there are several key functionality challenges to consider when PRO tools are brought to scale across the health system, highlighting the importance for continuous monitoring of questionnaire completion status, data documentation across care teams, and user feedback. Integrating process metrics and end-user feedback ensures that technical solutions are optimized to meet the information needs of diverse stakeholders as well as quality improvement and care delivery goals. Future work to understand how tailored EHR-enabled strategies can further reduce the time burden on both patients and their clinical care teams may facilitate improved response rates and adoption in practice. Assessing the implications for building PRO collection tools in the portal allowed us to forecast potential scalability limitations that would affect our ability to execute HRA and PHQ in the portal at the systems level. These findings provide foundational information for the next implementation phase to spread HRA to additional clinic sites and enhance PHQ functionality to support population-focused panel management.
National Prevalence and Hospital Characteristics Associated with Use of Health Information Exchange Following Discharge from Inpatient Psychiatric Units in Acute-Care Hospitals

Morgan Shields, M.Sc., M.A.,1 Grant A. Ritter, Ph.D.,1 Alisa B. Busch, M.D., M.S.2,3
1Heller School for Social Policy and Management, Brandeis University, Waltham, MA; 2McLean Hospital, Belmont, MA; 3Department of Health Care Policy, Harvard Medical School, Boston, MA.

Introduction
Psychiatric illnesses are prevalent and often associated with higher rates of co-occurring medical conditions.1,2 Acute-care hospitals with dedicated inpatient psychiatric units are important providers for patients with severe psychiatric illnesses: nearly 79% of US acute-care hospitals provide mental health services, and psychiatric units in them account for 71% of Medicare inpatient mental health discharges.3,4 New forms of treatment and payment models, such as accountable care organizations (ACOs) and integrated primary and mental health care, aim to better serve individuals with psychiatric illnesses.5–8 These new models will require robust use of electronic health information exchange (HIE) to manage transitions from inpatient to outpatient care.9 In 2016, 88% of U.S. non-federal acute-care hospitals electronically sent information to providers outside of their health system.10 However, HIE from inpatient psychiatric units in acute-care hospitals may not be as prevalent as it is from other inpatient units in these same hospitals. In part, this could be because psychiatric illnesses typically are accompanied by heightened privacy concerns from patients and more stringent federal and state privacy laws and regulations,9,11,12 which can influence how or whether health information is shared across providers. Despite the importance of HIE at discharge from inpatient psychiatric care in order to better manage severely ill and complex patients, no prior research has examined HIE to outside providers from psychiatric inpatient units in acute-care hospitals. Using national data, this study examines the reported use of HIE at transitions of care from acute-care hospital inpatient psychiatric units in 2016, and the hospital characteristics of these psychiatric units that are associated with HIE.

Methods
This is an observational cross-sectional study that merged facility-level data from 2016 the Centers for Medicare and Medicaid Services (CMS) Inpatient Psychiatric Facility Quality Reporting (IPFQR) program with data from the American Hospital Association (AHA) 2016 Annual Survey. The outcome of interest was whether an inpatient psychiatric unit within an acute-care hospital reported to the IPFQR whether it used an EHR to send HIE at times of transition in care. Logistic regression models were fit to examine characteristics associated with HIE from inpatient psychiatric units in acute-care hospitals (N=1,085, 99.8% of hospitals in the IPFQR; N=2 hospitals excluded because they were not in the AHA database of hospitals). Individual hospital characteristics in these models included overall hospital-bed size (<100, 100-399, ≥400), psychiatric-bed size determined by tertiles (3-8, 19-37, ≥37), teaching status (major teaching versus not), ownership (for-profit versus nonprofit and government versus nonprofit status), percent Medicaid bed days, urbanicity (metropolitan versus rural and micro areas), whether the hospital was part of a network, and whether it participated in an ACO. State level variables in the models included the overall state-level rate for electronically sending patient care summaries at discharge to any provider13 and a dichotomous variable denoting whether a state required patient authorization for sharing mental health information for treatment purposes.14 Multiple imputation was performed for item-level missingness on variables of HIE (12 missing), ACO (235 missing) and network (178 missing), with 5 iterations and assuming missing at random.

Results
In 2016, out of the N=1,073 acute-care hospital inpatient psychiatric units that responded to the HIE question in the IPFQR, 56.3% (N=604) reported they sent HIE from their EHRs to providers at the transition of care. Hospital characteristics associated with sending HIE from their inpatient psychiatric units were: large hospital bed-size (≥400 vs. <100: OR[95%CI]=2.46[1.46-4.15]) and hospital participation in an ACO (1.37[1.01-1.86]).

Conclusion
Sending electronic HIE from inpatient psychiatric units of acute-care hospitals lags far behind HIE use-rates that are happening more generally from the hospitals that “house” these psychiatric units (i.e., 56.3% vs. 88%). Hospital size

1457
and participation in ACOs—but not participation in a network—were associated with sending HIE from inpatient psychiatric units of acute-care hospitals. State mental health laws more stringent than HIPAA (that is, requiring patient authorization for sharing of mental health information for treatment purposes) were not associated with electronic HIE from inpatient psychiatric units of general hospitals. More research is needed to better understand characteristics associated with HIE from inpatient psychiatric units of acute-care hospitals. For example, whether outpatient provider EHR adoption rates or provider culture influence HIE from inpatient psychiatric units. Also, more research is needed to understand the potential role of alternative payment models to facilitate electronic HIE from inpatient psychiatric units in acute-care hospitals.

Acknowledgements

The authors gratefully acknowledge funding support from McLean Hospital (Busch), the Health Services Research Division of Partners Psychiatry and Mental Health (Busch), and NIDA (P30 DA035772).

References

Validation of a Computable Phenotype for Site-Specific Cancer Treatment

Evan T. Sholle, MS1, Orrin S. Belden, BS2, Jaclyn D. Rosenzweig, MD2, Jennifer Levine2, MD, Joseph Kabariti, BS1, Thomas R. Campion, Jr., PhD1,2,3, Jyotishman Pathak, PhD3
1Information Technologies and Services Department, Weill Cornell Medicine, New York, NY; 2Department of Pediatrics, Weill Cornell Medicine, New York, NY; 3Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY

Introduction

While widespread adoption of the electronic health record (EHR) has ushered in an era of unprecedented availability of patient data, challenges remain in secondary use of clinical data, including the translation of clinical concepts into concrete, machine-interpretable query algorithms represented using reference terminology and structured data points. To address this issue, the informatics research community has advanced the idea of a “computable phenotype” – a structured, machine-interpretable query algorithm that can be used for patient cohort identification using EHR data. Computable phenotyping has driven advances in the use of EHR data to study diverse conditions including diabetes, autism, and many others. Existing knowledge bases such as PheKB (1) also allow investigators to share locally-developed phenotype algorithms and make use of algorithms developed elsewhere.

While these efforts have yielded significant benefit towards defining patient cohorts, they have primarily centered on identifying phenotypes that relate solely to the patient’s physical state. However, at our institution, many patients who receive a structured diagnosis of cancer are receiving treatment for unrelated conditions and receive cancer care elsewhere. While these patients might meet the criteria for a “cancer” phenotype, their lack of cancer-specific data points would render them unsuitable for inclusion in research studies that depend on their receiving cancer care at our institution, as their EHR data would not contain any of the elements required to conduct retrospective research on oncologic outcomes. To remedy this issue, we sought to develop and validate a computable phenotype for a patient cohort defined not purely by clinical criteria, but also by healthcare-process-specific criteria: patients receiving cancer treatment at our institution, as distinguished from patients receiving cancer care at another institution but seeking care from our institution for other conditions.

Methods

We first developed a manually-annotated gold standard of patients who were receiving cancer treatment at our institution, Weill Cornell Medicine. This process included identifying a sample of 902 patients diagnosed with Hodgkin’s lymphoma at encounters when they were 15-39 years of age, as defined by the presence of an ICD-9 or 10 code anywhere in the electronic health record, including patient self-reported history, billing diagnoses, and problem list entries, as well as encounter diagnoses. We then annotated each patient with a binary classifier for either receiving or not receiving cancer care at our institution. Annotation was based on subjective clinical judgement, as we sought to determine the extent to which each phenotype criteria or combination thereof matched clinician adjudication.

After developing the gold standard, we applied three distinct criteria, along with their combinations (see Figure 1), to classify these patients as either receiving or not receiving cancer treatment at our institution.
The first criteria was whether or not the patient had 2+ encounters with an oncologist. The second was whether the patient had an order for a medication of the “antineoplastic” drug class as defined by the National Drug File - Reference Terminology (NDF-RT), and the third was an encounter diagnosis associated with an ICD-10 code for neoplasia. Using existing infrastructure for secondary use of electronic patient data (2), we then queried the EHR to determine whether each patient met each of these criteria. We compared the results of each criteria alone and in combination, to determine their performance relative to the gold standard by calculating precision, recall, and F-score for each.

**Results**

As detailed in Table 1, the various approaches yielded differing results, with the combination of all three criteria showing the highest specificity and the lowest recall. Overall, as determined by F-score, combining antineoplastic orders with either diagnoses or encounters demonstrated the best performance, with an F-score of 0.72.

**Table 1.** Precision, recall, and F-score for each combination of phenotype criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encounter diagnosis of neoplasia</td>
<td>0.35</td>
<td>0.98</td>
<td>0.51</td>
</tr>
<tr>
<td>Encounters with an oncologist</td>
<td>0.39</td>
<td>0.99</td>
<td>0.56</td>
</tr>
<tr>
<td>Order(s) for antineoplastics</td>
<td>0.74</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td>Diagnosis + encounters</td>
<td>0.42</td>
<td>0.98</td>
<td>0.59</td>
</tr>
<tr>
<td>Diagnosis + antineoplastics order(s)</td>
<td>0.76</td>
<td>0.67</td>
<td>0.72</td>
</tr>
<tr>
<td>Encounters + antineoplastics order(s)</td>
<td>0.78</td>
<td>0.68</td>
<td>0.72</td>
</tr>
<tr>
<td>Diagnosis + encounters + antineoplastics order</td>
<td>0.78</td>
<td>0.67</td>
<td>0.72</td>
</tr>
</tbody>
</table>

**Discussion**

The relatively poor performance of ICD-10 codes alone (as demonstrated in Table 1) as a proxy for ascertaining whether or not patients received site-specific treatment for a given condition is a major limiting factor in identifying patient cohorts for clinical trials, as well as retrospective research. Since patients with diagnosis codes alone may have the conditions in question, but did not receive care at the institution, they likely lack many of the data points required to support the research question for which their inclusion was determined. This often necessitates painstaking manual review to exclude and filter out patients from a cohort ostensibly matching study inclusion criteria. Similarly to the existing support biophysical computable phenotypes provide in identifying patients with given clinical characteristics, this technique may help researchers better characterize patients seeking care for a given condition at a specific institution, obviating the need for manual review to exclude patients without sufficient condition-specific data.

Healthcare fragmentation is a well-documented phenomenon (3), and it is possible that our relatively young cohort may be less subject to this issue than an older group of patients, artificially deflating the count of true negatives in our gold standard. However, we deemed the relatively higher availability of longitudinal cancer-related treatment in this younger, lower-mortality cohort to be a worthwhile tradeoff – future research may expand the utility of the classifier in older populations and other disease areas. Additionally, incorporation of data from collaborative research networks, such as New York City’s INSIGHT Network, may provide an opportunity to better characterize site-specific treatment.

**Acknowledgements**

This study received support from NewYork-Presbyterian Hospital and Weill Cornell Medical College, including the Clinical and Translational Science Center (CTSC) (UL1 TR000457) and Joint Clinical Trials Office (JCTO).

**References**

Sonoma County Empowering Caseworkers to Better Serve the Most Vulnerable

Jane L Snowdon PhD\(^1\), Barbie Robinson\(^2\), Carolyn Staats\(^2\), Kenneth Wolsey\(^3\), Thomas Strasheim\(^4\), William Kassler MD MPH\(^1\), Gretchen Jackson MD PhD\(^1\)

\(^1\)IBM Watson Health, Cambridge, MA, USA; \(^2\)County of Sonoma, CA, USA; \(^3\)IBM Global Business Services, San Diego, CA, USA; \(^4\)IBM Watson Health, Denver, CO, USA

Abstract

Sonoma County government agencies employed advanced health information technologies, artificial intelligence (AI), and inter-agency process improvements to transform health and healthcare for socially disadvantaged groups and other displaced individuals. This unique application of innovative care-management tools, which created a bridge between social and clinical determinants of health, helped 77 individuals who were at risk and experiencing a crisis through coordinated care plans and access to services in their time of need.

Introduction

State and local government agencies are faced with budgetary, organizational, and legislative constraints as well as increasing demand for services. Typically, agencies are separately funded, staffed, and operate in “silos of service” with limited inter-agency integration. Recognizing an unmet need for a better way to support their most vulnerable citizens, Sonoma County, California prioritized their safety net system in early 2017. While initiating the workplan to support improving their safety net system, the Sonoma County Complex fires hit burning more than 36,000 acres, killing 22 people, and destroying 5,636 homes and buildings.\(^1\)

Sonoma County has an estimated population of 500,000, is largely rural, and exhibits significant disparities in health, living standards, and educational attainment.\(^2\) Demographically, whites are the largest racial group. Latino or Hispanic persons represent the second largest and fastest growing group over the past decade with 27% of the total population. Sonoma County is challenged by a high cost of living, limited housing, and a homeless crisis that was exacerbated by the 2017 fires.\(^3\) In the aftermath of the wildfires, many vulnerable citizens already living in shelters needed ongoing shelter, while others who required shelter were faced with dwindling supply. Sonoma County sought to design and deliver better coordinated service delivery for its citizens, particularly for individuals experiencing one or more complex needs including unemployment, housing insecurity, homelessness, behavioral health issues, substance abuse, and incarceration. Sonoma County’s siloed departments (e.g., Health Services, Human Services, Community Development Commission, Child Support Services, Probation and others relating to criminal justice) collaborated through advanced care-management tools to decrease fragmentation of care for shared clients; reduce duplication of services; and create an effective, efficient care management plan.\(^4,5\)

Methods

Sonoma County implemented a three-step approach for achieving their objectives. First, a new program called Accessing Coordinated Care and Empowering Self Sufficiency (ACCESS) Sonoma County was established to identify the most vulnerable residents and match them with services to help improve their well-being and self-sufficiency, increase service utilization, reduce duplicate services, and lower expenditures per client or family. Second, Sonoma County created an Interdepartmental Multidisciplinary Team (IMDT) to generate and deploy coordinated cross-departmental services to address individuals experiencing homeless or housing insecurity needs and to inform the strategy of technology to support IMDT care coordination requirements. Third, a data integration hub (DIH) and care management and coordination system (CMCS) were implemented. The DIH brought together data from a variety of sources and created a single, integrated, current master data index resulting in a golden record for every individual. More than 91,000 resident records were loaded from across four sources (CalWIN for eligibility operations, SWITS for drug and alcohol dependency, Avatar for mental health, and IJS for proprietary justice data) into the DIH. The CMCS was built on a Health Insurance Portability and Accountability Act (HIPAA)-enabled, cloud-based platform that can aggregate data and connect agencies, payers, and community service providers, helping to support the design, coordination, and delivery of individualized care plans and services to

1461
manage the whole person. Across multi-disciplinary teams, care managers, social workers, and treatment providers optimized workflow for activities such as scheduling, developing individualized care plans, managing budgets, selecting providers, and enabling care. The CMCS uses one care plan to coordinate care across a multi-disciplinary team for a person’s mental health, physical health, and social service needs, such as palliative care, chronic disease management, personal budgets, and substance use management. The CMCS is able to apply best practices (e.g., management of comorbid conditions such as diabetes, heart failure, chronic obstructive pulmonary disease, hypertension) through structured programs built in alignment with evidence-based medicine, quality standards, and regulatory requirements. Individuals can engage with their care team and collaborate on their care plan through their own responsive application. Given client consent and local rules regarding the user’s role in the organization and eligibility to view the data, the CMCS enabled secure and systematized data collection, storage, and access. The CMCS provided the IMDT with a cloud-based 360° view of a client’s health, behavioral, and social factors and the ability for client-centric, goal-driven care planning with the aid of an AI component for automatically extracting and surfacing social and clinical determinants of health terms in large volumes of care management notes to practitioners.\textsuperscript{5-6} The CMCS allowed the IMDT to create and care manage a new fire-displaced cohort, perform assessments, record case notes, set care plans and actions, and match client needs with community providers.

Results

Caseworkers, clinicians, housing specialists, and other community partners coordinated services for the highest risk clients. More than 4,000 fire victims were initially housed in the shelter. After week two of the disaster, the number of individuals who did not have resources to leave the shelter for secure housing was about 300. Of 127 individuals representing 95 households that remained in shelters two weeks after the fires, approximately 42 percent were identified as precariously housed and approximately 26 percent were homeless prior to the fires. Over the ten-month period from April 2018 to February 2019, the IMDT managed 77 clients with multi-dimensional, complex needs. The median age was 54.5 years (range 29 to 84 years), and 58% were female. Care managers identified goals for each client (e.g., stabilize client by addressing emotional trauma and financial distress, rehouse client, and engage other collaborative care providers to address health and safety concerns). In some cases, care managers were unable to establish or maintain contact with the client. The average number of times IMDT formally discussed cases was 2.9. As of February 2019, 34 cases were stable and intermittently monitored and 43 were being actively managed.

Discussion

Care management tools have traditionally been applied in chronic disease or primary care settings. This innovative care-management solution transformed health and healthcare for socially disadvantaged groups and those displaced by a natural disaster. With the integrated DIH and CMCS solution, Sonoma County’s IMDT is now able to enter, aggregate and render near real-time information to caseworkers, clinicians, housing specialists, and other community providers. IMDT staff believe DIH and CMCS have resulted in productive care-planning meetings, information transparency, and enhanced access to coordinated care. With the assistance of DIH and CMCS, applications for food stamps, Section 8 housing, and Medi-Cal, California’s coverage for adults of limited means, were prepared and thoroughly reviewed by IMDT, reducing the burden on multiple staff. Advanced health information technologies can be instrumental in driving collaboration and access so that residents at risk and experiencing a crisis can be better served in their greatest times of need. Conceptually, the potential exists for replicating this program across other counties with possible operational efficiency and client outcome benefits.

References

3. Sonoma County and IBM working together to change lives: empowering caseworkers to better serve the most vulnerable. IBM Global Markets White Paper. February 2019.
Designing for the Information Needs of Primary Care Physicians

Jeffrey Sokolov, PhD, Morgan A. Foreman, BS, Yoonyoung Park, ScD, Sarah Miller, PhD, Amar K. Das MD, PhD, Paul Tang MD, MS

1IBM Watson Health, 2IBM Research, Cambridge Massachusetts

Introduction

At the point of care, clinicians seek information to understand their patients such as ‘What is this person’s HbA1C goal?’ or ‘Did this person have an adverse reaction to metformin?’. Today, finding answers to these questions often requires foraging through progress notes from past encounters in Electronic Health Records (EHRs). Our goal is to improve the clinician’s understanding of their patient by making it easier to find answers to their prototypical questions [1]. We used a previously developed artificial intelligence natural language processing (AI-NLP) method [2] to extract relevant information from unstructured data in clinician notes. We present a user-centered design approach to support the creation of a user interface for a primary care clinician at the point of care.

In the past, mental models have helped cognitive psychologists understand how people accomplish complex tasks in varied domains such as navigation [3], electricity [4] and others. We hypothesize that physicians’ mental models of what’s needed to understand and treat a specific person drive the search for relevant information in that patient’s medical record. This information will form the basis of what is needed to design a software product that displays a patient’s medical information where and when it is needed.

Method

A mixed methods approach was used to identify both the nature of physician’s prototypical questions as well as the order in which they are asked. The IRB approved research was performed in partnership with Atrius Health (a medical practice group in Massachusetts). A convenience sample of physicians was recruited from the clinical staff at Atrius Health. In Study 1, think aloud protocols were collected from eight primary care physicians at Atrius Health as they reviewed four patient records. For each record, we asked the physician to act as if she had inherited the patient from a retired physician and was about to see the patient for the first time for a chronic disease management follow-up. A total of twenty-seven different reviews were recorded, transcribed and coded using a grounded theory approach. For each unit of speech, the coding system captured both the information category (e.g., “Medications” or “Labs and Vitals”) as well as the intent behind the search (e.g., “Understand the patient” or “Formulate an hypothesis”). Study 2 was an observational study in which two authors (JS, SM) shadowed nine primary care physicians before, during and after fifteen different patient encounters at five Atrius Health locations. Observations were captured in handwritten notes with a special focus on the questions that physicians asked, either implicitly or explicitly. The notes were also analyzed for themes relating to mental models and clinical workflow. Study 3 was composed of a series of feedback sessions with clinicians where we tested the usability of the improved interface in answering the clinician’s prototypical questions and providing the right information in context. During these feedback sessions, clinicians at Atrius Health interacted with increasingly refined versions of a high fidelity prototype using the Invision software product. Continued refinement of the clinicians’ mental models and, most importantly, the display was made possible by triangulating the results of each study.

Results

From Study 1, we learned that all clinicians spent at least 50% of their time in a patient’s record searching for information to better understand the patient (range from 54-81% of all information seeking codes). Typically, they were searching for information about the patient’s Problem List, Labs and Vitals, Person Story and Medications. From Study 2, we were able to collect a broad set of questions that physicians ask about each of these information types. Study 3 enabled us to continually optimize the interface for usefulness and ease of use.

To illustrate one of the mental models, we present a record review of a patient focused on Type 2 Diabetes (Figure 1). This physician goes through a process of searching for information then formulating a grounded hypothesis, setting an agenda for the visit and then repeating that process to refine the agenda. The green boxes highlight areas where AI-NLP can be used to quickly pull information from the EHR to create a clinically relevant insight.

Application to Interface Design

After we analyzed the information seeking behavior and prototypical questions the clinicians asked, we used the mental models to develop a user interface. The goal of the display was to minimize the information foraging and time needed to find answers to the clinician’s prototypical questions.

To develop the interface, first we took the questions to determine what information the clinicians needed and where that information can be found in the EHR. Recall that information can come from both structured (e.g., labs, medications and vitals) as well as unstructured data (e.g., physician progress notes).
Type 2 Diabetes is a chronic condition, measured by HbA1C, that is often treated with medications as well as lifestyle factors. In addition to understanding the person’s HbA1C and whether it is controlled or not and the medications they’ve been prescribed to control their condition, clinicians often want to know about additional labs, the person’s weight and weight goals and the control of their blood pressure. The resulting interface visualized all of this information along a common timeline.

Three information-seeking behaviors illustrated in Figure 1 could be answered by insights extracted by the AI-NLP service from the text in the physician’s note. In each case, the extracted insights would be presented in context within the interface. For example, the PCP6 statement that “A lot of times when you [increase the metformin], you start developing diarrhea” would be captured as an “Adverse Drug Reaction” and presented as an insight associated with the display of metformin on the date that it was documented by the physician. In such a case, clinicians would be given a visual cue that there was an NLP-based insight and they could hover over the location to read the relevant insight (Figure 2). As confirmed from the feedback sessions in Study 3, this resulted in reduced foraging through the patient’s record to find relevant information.

Figure 2: The left AI-NLP insight is an example of information that would appear when a clinician hovered over the blood pressure goal on the Hypertension page, while the right image shows a Medication Change insight.

Acknowledgement

The authors wish to recognize and thank John Zambrano, MD and the many clinicians at Atrius Health who participated in this research.

References

Using Interactive Data Visualization to Drive Quality Improvement in Emergency Medical Services

Rachel A. Stemerman, MPH, NRP¹, Malvika Pillai¹
¹University of North Carolina, Carolina Health Informatics Program, Chapel Hill, NC

Introduction

In 2017, the National Emergency Medical Services Information System (NEMSIS) reported that 4,016 Emergency Medical Services (EMS) agencies responded to over 7 million EMS activations through the United States 911 system¹. EMS agencies vary significantly in the use of quality improvement programs due to resource limitations and data collection barriers². As a result, EMS providers may be unaware of opportunities to improve their care knowledge and skill. Measuring and comparing care quality against agreed-upon standards of practice or peer performance (audit) and providing results to healthcare professionals and or administrators (feedback) is a common quality improvement strategy³. Studies link audit and feedback to more accurate self-assessment of a health professionals’ performance³⁴. However, health professionals may overestimate their clinical performance and, in turn, fail to identify opportunities to improve their practice even if improvement is recommended. A number of studies have highlighted how dashboards had a positive effect on outcomes and may change individual provider behavior⁵. In this study, our objective was to develop and evaluate interactive quality improvement dashboards for EMS providers as guided by a user participatory design process.

Methods

We continued our work developing interactive visualization dashboards that displayed the EMS provider’s patient symptom assessment (or primary impression), their treatment intervention, and the resulting patient outcome. Scenario evaluations from the previous iteration found that these dashboards addressed system quality metrics, but not the individual needs of the EMS providers. Analysis of focus group data identified that burnout and lack of experience with specific populations such as pediatrics were of primary concern.

A user participatory iterative design process guided the development of two interactive dashboards created through D3.js. D3.js, an open source JavaScript library, was the primary software used and data analysis was performed with Python’s pandas library. One year of data was extracted from the Orange County Emergency Services (NC) EHR resulting in 300 - 450 patient contact per individual provider. Figure 1 shows a heat map calendar displaying the daily call volume for an individual EMS provider. Days with darker color indicate higher call volume. A tooltip on the individual calendar days displays the number of 911 calls the provider completed that day. When a day is clicked, a time series graph of their call volume appears at the bottom of the calendar (Figure 2) displaying the 911 call’s total length of time over their shift from 0600 until the next day at 0600 (y-axis).

The second dashboard displays all patient contacts the EMS provider had in 2018 as an interactive hierarchical bar chart (Figure 3). Each blue bar represents the total number of patient contacts categorized by the EHR data element “primary support system.” Clicking on a bar allows the user to drill down into the patient’s “primary support sign.” Another click on a bar populates the provider’s “primary impression” or their diagnosis of the patients’ active complaint condition. Clicking on the background will return the user to the previous bar chart until the home screen

Figure 1. Interactive heat map of EMS provider’s daily call volume

Figure 2. Time series line graph

Figure 3. Interactive hierarchical bar chart
is reached. This categorization allows the user to visualize the full scope of clinical patient contacts they evaluated that year highlighting low volume categories in red (< 10).

**Figure 3.** Hierarchical bar chart of clinical patient contacts for an individual EMS provider

Results

During a preliminary evaluation, all providers (N = 12) reported that this was their first exposure to their personal operational and clinical care data. As providers explored the calendar heat map, they expressed a better understanding of what was contributing to their feelings of burnout. Previously, providers perceived they had higher call volume, but were unaware of the total length of time of individual patient contacts. Evaluation of the interactive bar chart resulted in increased awareness of gaps in clinical patient contact types and a desire to pursue continuing education in areas of greatest need. Many providers also reported that, due to faulty memory, they underestimated the number of critical patient contacts (i.e., cardiac arrests) they experienced.

Discussion and Conclusion

We created visualizations that provided a method for EMS providers to rapidly and accurately analyze their individual operational and clinical care data resulting in actionable information. Interactive presentation of this information can engage users and generate new insights to inform training guidelines and address the quality improvement needs of the individual health professional. To the best of the researchers’ knowledge, no other studies were found using interactive data visualization in self-initiated quality improvement. Future work will explore how providers use and integrate the information provided by the dashboards into their decision making.

Acknowledgement: All authors funded by National Library of Medicine through grant number 1-T15-LM012500-01

References


Engaging with MSK Engage: A Digital Platform for Monitoring Patient Reported Outcomes in Oncology

Peter D. Stetson, MD, MA; Era Bani1,2; Kevin Shannon1; Naga Akella1; Olga Strachna1; Monica Allison1; Rita Reynolds1; Judy Hagerty-Paglia1; John Philip1; Wazim Narain, PhD1; Julia Fu1; Kenneth Seier1; Donna Boccamazzo1; Suzanne Kelson1; MaryAnn Connor, RN1; Rori Salvaggio, RN1; Michael Eubank1

1Memorial Sloan Kettering Cancer Center, NY, NY; 2Mt. Sinai, NY, NY

Introduction:
Collecting and monitoring patient-generated health data (PGHD) is associated with improved patient survival in oncology [1,2]. Many electronic health records (EHRs) and patient engagement vendors have begun to enable remote monitoring and patient-reported outcomes (PRO) data collection. However, open Informatics questions remain regarding how to standardize and visualize actionable PGHD for the care team [3], what integration touchpoints are needed with EHRs to support clinician workflow, how to titrate appropriate notifications to the care team for alarm conditions, how to avoid patient “survey fatigue”, and what form and degree of clinician engagement confers benefit to patients being monitored. We report here on MSK’s ongoing efforts to engage with patients to support oncology care, and to broaden and deepen our phenotyping efforts for translational research.

MSK Engage is an Informatics solution embedded within our patient portal for PGHD collection and monitoring. It was launched in 2016 coincident with opening our ambulatory surgery center, and it is now live enterprise-wide for both standard of care and research patients. It automates the distribution and collection of numerous types of survey instruments, including PROs, medical history, structured family history, patient preferences, and patient experience. It integrates with our scheduling system, our EHR, and portal secure messaging. Instruments and questions are housed in libraries and structured using the emerging FHIR questionnaire standard [4]. Trigger notifications can be set in its rules engine at the question level (e.g. “severe” PRO-CTCAE symptoms[5]), and multi-lingual support is enabled. The design is responsive for desktop and mobile use by patients. Our patient-run advisory council (PFACQ) participates actively in the design and naming of MSK Engage.

Our objectives in this operational evaluation were to 1) understand system adoption by both patients and clinicians, 2) develop and implement user-centered workflow integration for notifications, documentation and ordering, 3) understand and establish best practices for governance, 4) develop prototype visualizations of PGHD from n=1 to cohort-level displays, and 5) incorporate PGHD into our deep-phenotyping efforts for translational research.

Methods:
Descriptive statistics were used to assess systems usage and adoption. Ethnographic observation, focus groups, and user-centered design approaches were used to develop workflow integration touchpoints and visualization dashboards. Dashboards were developed in Tableau with underlying data residing within our clinical data warehouse.

Results:
The portal adoption is 81% among patients under active treatment. As shown in Figure 1, 31 instruments are now live in MSK Engage across 359 participating clinics. 36,516 unique patients have completed 98,616 surveys with a median completion time of 3 minutes. 74% of surveys are completed at home, as opposed to onsite by iPad. Patient completion rates are increasing, currently at an average of 50% for Q1 2019. Patient completion rates vary by instrument. Not all instruments are configured to trigger alerts (e.g. routine medical history forms), but those that do have triggered 6,835 alerts on 3,236 unique patients associated to 59 physicians and 202 staff, who escalated the alert 25% of the time. Of the 31 instruments deployed to-date, 22 are standard of care, 2 are for research, and 7 are for both. 13 have a dedicated Tableau dashboard live or in development, 5 are imported directly into electronic documents in the EHR, and 1 generates automated referral orders (triggered when nurses review with the patient and save their linked note). Promotion to the legal health record, as a proxy for “provider attestation”, varied by instrument. An average of 2.7 instruments have been submitted per patient, enabling multiple time-point assessments for many patients.

Figure 2 demonstrates a user-centered prototype for cohort-level display of data in Tableau. Patient-level data views were developed as drill-downs. Additional enterprise dashboards have been developed in Tableau to monitor symptom alert triage and clinic compliance rates. At the individual survey response level, “advanced reports” have been developed for clinicians for 7 of the instruments, and for patients for 2 of the instruments.
To govern this process, we established an eForms Committee with multidisciplinary membership. It is staffed and run by Informatics, meeting twice monthly (working group; alternating with full committee). The committee is charged with avoiding duplicate questions and survey fatigue, as well as standardization of questions in the question library and triaging requests for alert triggers, documentation integration, and dashboarding requests. Patients who used the system generated new feature requests - most interestingly to add the capability to “ad hoc” report on symptoms even after the target cohort period was complete. An additional 15,000 structured family histories have been backloaded into Engage, with go-forward collection through Engage using the Family History Questionnaire for targeted research patients.

Discussion:
Patients and staff are adopting the platform for both standard of care and research, with increasing patient compliance. Patient and clinicians are using all the designed workflow integration tools, including direct import into documentation templates within our EHR, push notifications, and decision support. There is extensive and timely engagement by our nursing staff with MSK Engage data and triggered notifications on alarm symptoms. Opportunities exist to enhance physician engagement with the data. Significant oversight and governance with advanced visualization tools are required for success. MSK Engage has contributed multiple surveys per patient over time, which augments our clinical and administrative data sets for our ongoing phenotyping efforts. Future directions include enhancing the point-of-care display of this data for the clinical team, introducing tailored education information based on symptoms, and elicitation of patient treatment preferences.

Figures & Tables:
Figure 1: MSK Engage Overview

References:
Linear Algebra and Public Health: Identifying Areas of High Healthcare Needs through the Markov Chain Ranking Method

Abdul A. Tariq, PhD, Beatriz Ryan-Lora, MPH, Andres Nieto, MPA, David K. Vawdrey, PhD
1NewYork-Presbyterian Hospital, New York, NY, USA

Introduction

In recent years, increasingly larger quantities of data on the demographic, health, environmental and socioeconomic characteristics of communities are becoming available. While such data can paint a rich portrait of a community, it is unclear how to aggregate such disparate information into a single metric reflecting the overall “health” of a community. In this study we apply the Markov Method to aggregate multiple characteristics of a community into an overall fitness score. We apply the method to Census-tract level data on disease prevalence and health behaviors using a dataset from the Centers for Disease Control (CDC), to generate a healthiness ranking for areas in Manhattan: a large, ethnically and economically diverse borough in the city of New York. This score can be used by to pin-point areas where public health investments should be prioritized. The method is flexible enough to incorporate a variety of data types, and to build customized rankings based on different objectives.

Methods

The Markov Method is used extensively in rating and ranking, and among its most famous applications is PageRank, one of the earliest algorithms that powered Google’s search. We adapt it to our context and assume that we have \( i = 1 \cdots N \) communities, and each community has \( k = 1 \cdots M \) characteristics. For each of the community characteristics \( k \), we compute an \( N \times N \) voting matrix \( V^k \). In the voting matrix, each community casts a “vote” for every community that is better than itself (and there are \( N \times N \) such comparisons) on a given characteristic. In particular, the \( i \)-th entry of \( V^k \) is defined as follows: on a given characteristic \( k \) if community \( i \) is better than community \( j \), then \( V^k_{ij} = 0 \) (\( i \) does not cast a vote for \( j \)); or if \( i = j \). If community \( i \) is worse than community \( j \), then \( V^k_{ij} = 1 \) (\( i \) casts a vote in favor of \( j \)). If both are statistically indistinguishable (using some suitable test; we use the t-test), then \( V^k_{ij} = 0.5 \) (each community gets half a vote). We then row-normalize the voting matrix (each element is divided by the sum of all elements in that row). For rows having all zero elements (i.e. a community that beats all others on a given characteristic), we replace each entry with \( 1/N \). The resulting matrix is a stochastic matrix, and intuitively it reflects the probability of community \( i \) beating community \( j \) on a given characteristic.

Following the procedure above, we build a voting matrix for each of the \( M \) characteristics, and compute a final stochastic voting matrix as:

\[
V = \alpha_1 V^1 + \alpha_2 V^2 + \cdots + \alpha_M V^M
\]

Here \( \alpha_k \) is the weight assigned to the \( k \)-th characteristic, and all weights sum to 1. We then solve the eigensystem \( \lambda \mathbf{r} = \mathbf{r} \), where \( \mathbf{r} \) is the dominant eigenvector of the stationary distribution represented by the stochastic voting matrix \( V \). The \( i \)-th element of the dominant eigenvector represents the probability that community \( i \) will be better than all other communities, and we can use these values to compare communities with each other.

As an application, we rank different geographies within Manhattan on set of 23 health-related characteristics. We use census-tract level data from the CDC’s 500 Cities project, which provides the prevalence of a variety of chronic conditions and health-related behaviors for cities across the U.S. The 23 measures are bucketed into three overarching categories: health outcomes, unhealthy behaviors, and prevention.

Results

In Figure 1 we geo-code the rankings for the 2117 census-tracts in Manhattan for which data was available from the CDC. The rankings go from 0 (best) to 100 (worst), and are color-coded so that more red implies a worse score. The 23 health measures which comprise the ranking are provided in the bottom-left of the figure. To construct the ranking we use the crude prevalence of each of the 23 measures, as provided by the CDC. From examining the figures, two main points emerge: first, there is quite a bit of spatial heterogeneity in the overall health rank of areas within Manhattan. Even geographically contiguous areas can have vastly different health ranks, which underscores the need...
for geographically targeted public health interventions. Second, even the communities which are highly ranked overall, or in terms of health outcome measures, can sometimes fare poorly in terms of prevention measures. This highlights a potential disadvantage of aggregating multiple metrics into one (either through Markov or any other method), namely that it can potentially obfuscate areas in need.

**Conclusion**

We demonstrate a new method to aggregate a variety of community characteristics into a single ranking. Such a ranking can be used to prioritize areas in need for public health interventions, as well as provide an easily comparable summary of disparities across geographies.

**References**

Progressive Multi Class Predictive Modeling of Disposition Decision at Different Care Delivery Stages in an Emergency Department

M. Samie Tootooni, Ph.D. 1, Kalyan S. Pasupathy, Ph.D. 1, Heather A. Heaton, M.D. 1, Casey M. Clements, M.D., Ph.D. 1, Mustafa Y. Sir, PhD. 1  
1Mayo Clinic, Rochester, Minnesota, USA

Introduction

Boarding time in Emergency Department (ED) is defined as the time from the admission decision for a patient until the patient is placed in an inpatient bed. Higher boarding time is shown to be associated with many adverse outcomes such as mortality 1. Early prediction of the disposition decision, specifically potential inpatient admission, can facilitate timely initiation of the hospitalization process and therefore shorten the boarding time 2.

At the time that a patient arrives to an ED, only limited amount of information such as modality of transport, gender, and date of birth are usually available to the ED care team. While the patient receives care through different stages, the ED care team gets progressively more information about patient and consequently, prediction of disposition decision becomes increasingly accurate with the accrual of available electronic health records (EHR) data. Accordingly, we sought to develop, validate, and compare different multiclass classification models to predict the ED disposition decision (i.e., discharge home, admit to hospital, admit to intensive care unit, or ICU) for each patient at each major care delivery stage.

Recorded in the earliest stage of a patient’s visit, chief complaint affects several facets of a patient’s ED visit, i.e. nursing focus, initial medical assessment, prioritizing the treatment, determination of patient flow through the ED, and generation of different hypotheses for the final diagnosis 3. We recently developed a natural language processing approach (so-called CCMapper) to map these free-text data into a structured list so that they can be used in predictive studies. We evaluated whether incorporating chief complaint category improve the performance of the prediction models.

Methods

Three ED care delivery stages for visitors are identified: (1) patient arrival, when the registration clerk records the patient’s demographic data and chief complaint; (2) patient triage, when a nurse measures the height, weight, and other vital signs to determine emergency severity index (ESI level); and (3) first order bundle, when the first set of orders (medication, lab tests, etc.) is submitted by an ED Physician. We consider the first care delivery stage as time zero (T1 = 0). Table 1 shows the distribution for time to each care delivery stage as well as the disposition time.

Table 1. The distribution of time to care delivery stages.

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>Mean</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival to Triage (T2)</td>
<td>34</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>36</td>
<td>118</td>
</tr>
<tr>
<td>Arrival to First Order (T3)</td>
<td>40</td>
<td>3</td>
<td>10</td>
<td>22</td>
<td>47</td>
<td>100</td>
</tr>
<tr>
<td>Arrival to Disposition</td>
<td>436</td>
<td>166</td>
<td>247</td>
<td>362</td>
<td>509</td>
<td>685</td>
</tr>
</tbody>
</table>

The study cohort included all adults who visited the Mayo Clinic ED in Rochester, MN, between January 1st 2016 and December 31st 2017 (~140,000 encounters). In order to predict the final disposition decision, we trained, validated, and compared three types of predictive models: Logistic Regression (LR), Random Forest (RF), and Naïve Bayes (NB). All models used 10-fold cross validation with 21 months of data reserved for training the models and the remaining 3 months used for testing. By default all instances in the training data has equal weight for learning algorithm. To minimize the effect of unbalanced classes, we reweight the instances in a way that sum of weights for all classes of instances in the training data are equal.

Results
At each care delivery stage, we identified the significant variables that have high predictive power regarding the disposition decision. We found that patient transport categories such as *air ambulance* and *surface ambulance* are associated with higher risk of ICU admission while other categories, particularly *walk-in* is associated with lower ICU admission risk. Several chief complaint categories such as “accidental overdose” “blood pressure” and “rule out cerebrovascular accident” were found to be associated with ICU admission; in contrast with categories such as “lower/higher extremity pain” or “ear pain,” which had a negative relationship with ICU admission risk. Our analysis also showed higher age and male sex are associated with ICU admission.

Several patient factors associated with ICU admission are identified at the triage and first-order-bundle stages. For instance, *lower SPO2*, *higher acuity level (ESI)*, and ordering *lab-work, EKG, Radiology, CVT, etc.* are associated with ICU admission. Ordering medications such as *sodium, levofloxacin, vancomycin, piperacillin, and etomidate* are associated with higher ICU admission, in contrast with medications such as *ibuprofen, ketorolac, and lidocaine*.

All tested models had similar performance (e.g. predicting ICU admission at stage 3 (T3) with LR showed an Area Under the Receiver Operating Curve (AUROC) of 0.90 where RF and NB had AUROC of 0.89 and 0.86 respectively). However, LR slightly outperformed the other models in the majority of cases. Figure 1 demonstrates improvement of AUROC of the LR model from stage 1 (T1) to stage 3 (T3) on the test set. Also applying CCMapper and using its corresponding structured list of chief complaints categories not only provides insight about the general categorization of ED visit, but also improves the performance of the prediction models in all care delivery stages compared to not considering the chief complaint, or just consideration of the most frequent ones.

**Figure 1.** The AUROC of the developed progressive Logistic Regression model at each care delivery stage.

**Discussion**

The developed progressive predictive model uses the available information at each care delivery stage to predict the disposition decision. Barak-Corren et al. recently developed a binary class progressive logistic regression model to predict hospitalization in three defined timestamps in ED: within the first 10 min, within an hour and within 2 hours. We aim to predict disposition decision in three defined care delivery stages instead of time-since-arrival [due to the available EHR data may considerably vary between patients in similar time-since-arrivals].

The developed model can predict the disposition decision at earlier stage of patient ED visit with high performance. This model has potential to be adapted in ED settings in order to reduce the boarding time and eventually to improve the patient flow. Also certain patient-specific data such as clinical notes and system information data such as number of available physicians or number of patients in the ED were not considered in the current version of the model. These additional variables can improve the accuracy of the models reported above.

**References**

A Data-Driven Approach for Integrating Infant Mortality Prediction and Survival Time Estimate from Electronic Health Records

Fuchiang (Rich)Tsui, PhD1,2,3, Luge Yang, PhD1,3, Lingyun Shi, MS1, Lezhou Wu, PhD1

1Tsui Laboratory, Children’s Hospital of Philadelphia, Philadelphia, PA, USA;
2Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA;
3Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA

Abstract

We aimed to predict infant mortality and survival time from electronic health records. We collected 11 years of delivery data from one women’s hospital and birth/death certificates from a county office. We developed data-driven predictive models that provided mortality risk and survival time. Their AUCs ranged 0.88-0.94. To our knowledge, we are the pioneers merging predictive modeling and survival analysis. Our models can potentially identify moms or infants with high-risk infant mortality in practice.

Keywords: infant mortality, machine learning, survival analysis

Introduction

There is an unmet need of predictive models for identifying moms or infants with infant mortality risk. Infant mortality rates in the U.S. in general have gradually decreased from 6.84 in 2003 to 5.87 in 2016 per 1000 births1. However, the U.S. is ranked 55th for infant mortality given our highest healthcare expenditure in the world2-3. Current state-of-art approaches for infant mortality prediction leveraged available EHRs. These models’ performance measured by the area under the ROC curve (AUC) ranged from 0.62 to 0.945. Moreover, survival analysis is commonly applied in clinical trials but rarely applied to mortality prediction. Similarly, machine-learning based predictive models rarely provide individual survival time estimate.

In this study, we hypothesized that we can develop a data-driven predictive model that can predict infant mortality and survival time. The contribution of the study includes multiple predictive models with critical risk factors that impact infant mortality and survival time prediction based on the risk factors in a predictive model.

Methods

This section describes study datasets, predictive modeling, survival analysis, evaluation metrics and validation approach.

Datasets

We collected two datasets: 1) the Magee Obstetric Medical and Infant (MOMI) dataset from the UPMC Magee Women’s Hospital and 2) the death and birth certificates from the Allegheny County Department of Human Services (ADHS). The study period is between 2003 and 2013. MOMI dataset is a curated electronic health record dataset comprising 266 variables for all babies and moms during deliveries at the hospital.

Predictive modeling

We developed six machine learning methods: Naïve Bayes (NB), K2 Bayesian network (K2), Ridge regression (RR), elastic net (EN), gradient boosting machine (GBM), and random forest (RF). We developed a greedy search data-driven modeling approach from feature selection.

Survival analysis

We developed a Proportional Hazards (PH) model6, a classic method to study the relative risks of death, which is the foundation for an infant’s survival time estimate. Instead of using conventional approach of PH model building that may not be optimized for infant mortality prediction accuracy, we used the risk factors identified from our predictive model to build the PH model, \( h(t|X) = h_0(t) \exp(\beta_1 x_1 + \cdots + \beta_p x_p) \), where \( h(t|X) \) represents a PH model at time \( t \), \( X \) is a vector with \( p \) variables (features) identified from the ridge regression model, and \( h_0(t) \) is a baseline hazard function independent from the vector \( X \).

Evaluation

For predictive modeling, we used the area under the ROC curve (AUC) to evaluate predictive models’ performance. We used 10-fold cross validation to evaluate each model. For survival analysis, we used three metrics: concordance, integrated AUC (iAUC), and Kaplan–Meier weighted Brier score. Concordance measures the probability of agreement between predicted survival and observed survival7. iAUC measures AUC at time \( t \). Since infant death is a rare event, we selected Kaplan–Meier weight function6, a common function to adjust the case-control imbalance for Brier score. We used 5-fold cross validation to evaluate the PH model. We used restricted mean survival days for survival time estimate.

Results

We describes predictive models performance and survival analysis results followed by sample case outcomes.

1473
**Predictive models performance** After following data quality assessment (completeness, validity, consistency, and currency), we identified 102 variables for predictive and survival modeling. We identified 495 singleton infant deaths (infant died within one year of life) from the two linked datasets and 75,842 singleton births (controls; survived within the first year of life). Table 1 summarizes the performance evaluation for the 6 models with the number of features used in each model. Gradient boosting machine had the best prediction performance with 102 variables. Ridge regression had similar performance with only 29 variables. The 29 variables in the Ridge regression model include several categories: mom demographics, mom behavior, healthcare utilization, infant/mom diseases and symptoms.

Table 1. Predictive Model Performance. Each cell represents the performance of each algorithm measured by an area under the ROC curve (AUC) followed by a number of variables (features) in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>NB</th>
<th>K2</th>
<th>RR</th>
<th>EN</th>
<th>GBM</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>93.28% (38)</td>
<td>93.28% (38)</td>
<td>93.33% (29)</td>
<td>88.32% (29)</td>
<td>93.6% (102)</td>
<td>92.2% (102)</td>
</tr>
</tbody>
</table>

**Survival model performance** The PH model with the 29 variables identified from the Ridge regression model had 92.29% of concordance, 93.76% of iAUC, and 0.0035 of Kaplan–Meier weighted Brier score. To illustrate the integrated mortality risk prediction and survival time, we selected three sample study cases. Case 1 has risk factors including gestational age < 23.5 weeks, birth weight < 2500g, and African American mom. Case 2 has all the Case 1 risks with maternal narcotic addition. Case 3 has all Case 2 risk factors with mom’s high school or lower education.

Table 2. Sampled predictive model output with both mortality risk estimates and survival days.

<table>
<thead>
<tr>
<th></th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(infant mortality</td>
<td>Data)</td>
<td>53.87%</td>
<td>83.36%</td>
</tr>
<tr>
<td>Restricted mean survival days</td>
<td>133</td>
<td>76</td>
<td>55</td>
</tr>
</tbody>
</table>

Discussion and Limitations To our knowledge, we are the pioneers developing a predictive model that can provide both risk level and survival time as shown in Table 2. We built the survival model based on the risk factors (variables) used by a predictive model tuned for best risk prediction. Such integrated modeling not only can serve a clinical tool for clinicians and social workers to prioritize intervention and resources based on survival time but also an education tool for educating potential population in risk to better understand the impact of infant mortality given mom’s behavior. We chose Ridge model for building PH model since it only used 29 variables (instead of 102) with similar performance. The PH model demonstrated the capability of survival time estimate.

The limitations of this study include 1) the results were based on a single hospital data; 2) the MOMI dataset provides curated data at the time of delivery, which may limit the timeliness of risk estimate.

Future Work We are developing additional predictive models based on high impact symptoms/diseases identified from the Ridge model such as pre-term delivery for early prediction of mortality risk.

Conclusions In this study, we integrated machine-learning based predictive modeling with survival time estimate. We also developed six infant mortality predictive models. To our knowledge, we are pioneers in the innovative work.

Acknowledgements The study was supported by the Richard King Mellon Foundation.

References

Organizational Diagnostics: A Systematic Approach to Identifying Technology and Workflow Issues in Clinical Settings
Kim M. Unertl, PhD, Laurie Novak, PhD, Joyce Harris, MA, Christopher Simpson, MA, Nancy Lorenzi, PhD
Vanderbilt University Medical Center, Nashville, TN

Introduction
Healthcare organizations are under enormous pressure to improve ambulatory access, reduce expenses and optimize operational processes. Achieving peak efficiency and effectiveness is challenging, as healthcare organizations are highly complex sociotechnical systems. Making changes to optimize functionality in one area can lead to unintended negative consequences in other areas. As new health information technology (IT) comes into an organization, conflict can arise between out-of-date practices and process requirements related to new technology. An evaluation of current workflow and operational practices is critical to ensure that health IT can be smoothly integrated. However, there is limited research addressing how to rapidly diagnose organizational issues generated by the interaction between health IT and work practices in a structured manner and to identify solutions. Based on this methodology gap, the People, Organizations, and Technology Research Group at Vanderbilt University Medical Center (VUMC) developed an integrated set of assessment methods known as Systematic Iterative Organizational Diagnostics (SIOD). The purpose of this podium abstract is to report on the development of SIOD, to describe the outcomes of its initial application, and to explore how this approach can be applied in other organizations.

Design Process for Systematic Iterative Organizational Diagnostics (SIOD)
The goals of developing the SIOD method were to provide rapid diagnoses of operational issues with an emphasis on health IT and to identify evidence-based pathways for improvement. To accomplish these goals, the team developed a comprehensive diagnostic method combining social science methods with data analytics. SIOD consists of five main stages: Background Scan, Engagement Building, Data Acquisition, Data Analysis, and Reporting. The Background Scan stage focuses on collecting baseline data about the clinical environment (e.g., staffing model, patient load, throughput) and conducting open-ended key informant interviews to build rapport and identify contextual assessment goals. During the Engagement Building stage, the research group engages with the entire healthcare team by holding a series of design workshops based on concepts developed by W. Brendan Reddy to understand environmental and contextual factors. In the Data Acquisition stage, qualitative and quantitative methods are used to collect detailed data about clinic workflow, spatial design, and other contextual factors. In the Data Analysis stage, qualitative and quantitative data are initially analyzed separately, and then combined to identify gaps in the diagnostic process and to triangulate patterns across data sources. During the initial development of SIOD, qualitative data were analyzed using an iterative open-ended approach, which led to development of a codebook for application across data sets. As part of the Data Analysis stage, the team begins to identify potential improvements/solutions for issues identified in the data. These initial improvements/solutions are compiled and used in the final SIOD stage, Reporting, which involves reporting on our findings to multiple stakeholder groups. Reporting sessions provide a form of member checking, where the assessment team verifies data analysis with people from the clinical site. Major stakeholder groups include: the sponsoring organizational management representatives, clinical site leadership, and staff/providers from the clinical site. The Reporting stage serves three purposes: handing off recommendations to leaders with the experience and resources to implement change, determining which recommendations organizational and site leadership want to pursue, and sharing cross-organizational issues.

Internal VUMC SIOD Applications
As of March 2019, the POTR group has applied the SIOD approach in two ambulatory clinics and one multidisciplinary clinical center. In addition, components of the SIOD process were used during the group’s longitudinal evaluation of an organization-wide large-scale Epic implementation. Throughout all applications of the SIOD process, our team has focused on identifying what aspects of the methodology delivers the most benefits, determining areas for methodology improvement, and refining data collection instruments and data analysis process.

Over the course of applying the SIOD approach across multiple contexts, we have developed a coding schema grounded in the data, used for qualitative data analysis and integrated into presentations at the end of SIOD assessments. The schema includes five major categories, grounded in the data: Putting patients first, Reducing the chaos, Matching space to function, Technology making work harder, and Staffing is more than numbers. Putting Patients First focuses on the connections and gaps between the concept of patient-centered care and day-to-day
requirements of clinical work. **Reducing the Chaos** focuses on the concept that although adaptability and organizational resilience are strengths, rapidly changing contextual factors (e.g., new health IT features, new processes for collecting information) can lead to inefficient and confusing workflow. **Matching Space to Function** recognizes the fact that space, even in a large academic medical center, is a finite resource and that healthcare organizations need to optimize how available space is used. **Technology Making Work Harder** focuses on how technology or the implementation of technology increases difficulty in clinical work. Although health IT has potential to make clinical work easier and build connections across healthcare teams, in some cases health IT can increase workloads, disrupt work practices, and introduce communication challenges. Finally, **Staffing is More Than Numbers** focuses on the idea that staffing is not just about the number of people working in a clinical site, but rather how staff are deployed (e.g., scheduling), how staff roles are defined, and how satisfied staff are with the working environment.

For each SIOD application, we have identified multiple recommendations for improvements, which our team hands off to clinical and organizational leadership. Recommendations span a variety of topics, with some recommendations being clinic-specific and some recommendations potentially for application across the organization. Our data analysis process also identifies cascading chains of recommendations, where acting on a higher-level proposal would also resolve multiple smaller issues. Each recommendation is mapped onto a 2x2 grid with axes of implementation difficulty and potential benefit, allowing leadership to prioritize interventions aligned with available resources while maximizing benefit.

**Implications for Replication of the SIOD Process in Other Organizations**

Healthcare organizations spend millions of dollars implementing health IT systems, but without alignment to organizational requirements and integration into clinical workflow, these investments fail to deliver the intended benefits. The SIOD process connects academic biomedical informatics researchers embedded within the organization to organizational diagnostic needs. Such academic-operational integrations hold promise for future management projects that are evidence-based, data-driven, and integrated with health IT. Several key drivers of success were identified during application of SIOD to date, and are areas needed to transfer this approach to other organizations. Involvement of organizational leadership extended throughout the process, from identifying clinical sites in need of rapid diagnostics to recommendations hand-off. Identifying approaches to get everyone on the healthcare team involved in providing feedback was critical, given the holistic view of clinical functionality embedded in the SIOD process. Issues are not just related to a single role but to the full clinical ecosystem, including the health IT. The recommendation hand-off portion of the Reporting phase was especially important, as multiple clinical staff members discussed frustration with past process improvement efforts where they had provided feedback but no action was taken. Implementation of recommendations was beyond the scope of SIOD, which served primarily to diagnose issues and provide change recommendations, but we had to hand those recommendations off to people with the ability to take action. Understanding barriers to rapid assessment Conducting assessments in an organizational setting requires substantial coordination of activity and substantial time and effort from personnel. Proactively identifying assessment team members, providing training on data collection and analysis methods, and ensuring adequate resources for administrative aspects of SIOD implementation are all key components to overcoming barriers to applying SIOD to other clinical contexts.

**Conclusion**

Our team developed a rapid methodology to diagnose organizational issues generated by the interaction between health IT and work practices. After applying the SIOD approach multiple times in ambulatory clinical settings, we have identified common patterns related to technology-work mismatches, need for data-driven decisions on contextual factors such as staffing and space allocation, and gaps between the goals of patient-centered care and the requirements of day-to-day work. We are continuing to develop and apply SIOD to new clinical environments and are working to establish a toolset for use across organizations.

**References**

Expanding the Representation of Permissions and Deontic Roles in the Informed Consent Ontology (ICO)

Jonathan M. Vajda, MA1, J. Neil Otte, PhD2, Cooper Stansbury, MS3, Marcelline R. Harris, PhD3, Elizabeth Umberfield, BSN, RN3, Frank J. Manion, PhD3, Cui Tao, PhD4
1University at Buffalo (SUNY), Buffalo, NY; 2Johns Hopkins University Applied Physics Laboratory, Laurel, MD; 3University of Michigan, Ann Arbor, MI; 4University of Texas Health Science Center at Houston, Houston, TX

Introduction

The Informed Consent Ontology (ICO) is an OWL ontology that aims to represent the domain of informed consent, including the various processes involved in consent. Consent includes, for example, informing participants of the key elements of a research study and what their participation will involve, and ensuring that a patient’s participation is voluntary. The Informed Consent Ontology also represents the complex set of regulatory, legal, and policy processes involved in regulated research, and represents the permissions that may be granted to a research team and various governing bodies (including funders, institutions, and collaborating groups) as a result of the consent process, as well as the realization of those permissions in permitted actions. The specific use case driving the expansion of ICO is the sharing of biospecimens and associated data, especially in biobank repositories. This present work addresses a critical gap in representing permissions robustly, by enumerating deontic roles serviceable to different organizations across informed consent contexts, and by incorporating the Common Rule while considering use outside the United States.

ICO relies on the Basic Formal Ontology (BFO), imports the Information Artifact Ontology (IAO) and the Document Acts Ontology (D-Acts), and can extend the Ontology for Biobanking (OBIB) and the Data Use Ontology (DUO). BFO is a realist, upper-level ontology that has been successfully employed in hundreds of projects, including those ontologies of the Open-Biomedical Ontology (OBO) Foundry. Since BFO is a true formal ontology containing only high-level classes like ‘process’, ‘quality’, and ‘material entity’, it is suitable for use across domains and facilitates semantic interoperability with other ontologies that import it. The IAO is a mid-level ontology that imports BFO and represents information entities such as documents, data, and plan specifications. The D-Acts Ontology is a small mid-level ontology that imports the IAO and represents document acts (e.g. creating obligations through documents, such as a contract) and deontic roles – a role that inheres in an agent and which is externally grounded in normative expectations concerning how that agent should behave. OBIB extends the Ontology for Biomedical Investigations (OBI), and represents annotation terms relevant to biorepositories. DUO uses representations of consent codes to annotate data with usage permissions. The Common Rule Ontology (CRO), a related informed consent ontology project, represents the US Common Rule.

Methods

Developers performed a refinement and restructuring task known in the software development community as ‘refactoring’; our process involved exporting the label and definition of each resource in the existing Informed Consent Ontology into a spreadsheet and marking resources for revision and deprecation. Complex use cases—thirteen in total—were developed to assess whether classes in ICO were necessary and whether new classes were required to improve expressivity. Literature on informed consent was consulted, including regulations and best practices for obtaining patient consent and consent form creation. Blank consent forms with a diversity of presentation styles, participation descriptions, and procedure directives were annotated with ICO classes to demonstrate coverage and expressivity at different levels of granularity. Resources were marked for revision when they represented entities that were necessary but could be more appropriately labeled or defined. This often involved making classes more atomic in structure. Resources were deprecated if they fell outside the intended scope of a reference ontology for informed consent or did not adhere to BFO development principles (e.g., ill-formed genus or species terms, such as the process ‘explaining to participant about research investigation’). OWL reasoners, such as HermiT, checked for consistency throughout development. We compared the resulting resources in ICO against the thirteen use cases using diagrams expressing the complex relations involved and representation available in the ontology (e.g., a query for directives permitting research on samples from persons diagnosed with Parkinson’s disease). Collaborators from other ontologies (D-ACTS, OBIB, and DUO) provided feedback and corrections on the representation’s adequacy for use by application ontologies. CRO’s classes were deprecated, retained, or revised (often through generalizing a class), and then integrated into ICO. These updates were made with a view toward international regulatory contexts.
Findings

We expanded the representation of the most general entities in tandem with the imported ontologies. ICO originally asserted 669 classes and CRO asserted 1244; the new ICO asserts 1125 classes. Of those classes retained, 100 labels and 150 definitions were refined. We improved the representation of informed consent forms across phases of completion, and recast the relationships between documents and their parts. We elucidated how directives prescribe deontic roles, such as ‘deontic power role’ by which one may create or revoke permissions. For example, ICO: ‘description of risk’ is an IAO: ‘directive information entity’ which prescribes certain individuals and institutions to perform specific action types despite the attending risks. Likewise, a data field that directs the user to check ‘Yes’ or ‘No’ for a certain statement becomes a new directive which prescribes some action must (or must not) be taken. The class ‘stasis of regulation’ (a subclass of BFO:occurent) enables tracking whether, where, and when certain policies and directives are in force. This overcomes a limitation in OWL, since a ‘stasis’ can be used to indicate time, place, etc. The use of ICO’s class IRIs for tagging a variety of informed consent forms (research and clinical forms) showed ICO’s coverage for the core parts of the forms, including descriptions, and directives, according to expected needs of users, as elicited by domain experts.

Discussion

With the new release, we have enabled ICO to be more useful as a mid-level reference ontology, importable into an application ontology tailored to a specific research or practice area, using a BFO-based ontology-driven information system (ODIS). Permissions can be represented as an ‘information directive entity’ prescribing a ‘deontic role’. One common complaint of OWL, the most commonly used ontology language, is the lack of time indexing for when agents participate in processes. By introducing the ‘stasis of regulation’ class, we provide an important and new contribution to the BFO-aligned ecosystem. The new release provides further granularity to differentiate informed consent processes which produce a legally effective consent from those who do not. Finally, the revised ontology avoids design issues in asserting negative claims where information is simply unknown (a violation of open-world assumption).

Acknowledgements

This work was supported in-part by the National Human Genome Research Institute of the National Institutes of Health under award number 5U01HG99454 (Tao, PI), and a Michigan Institute for Data Science’s (MIDAS) Challenge Award (Harris, co-PI). The content does not necessarily represent the official views of the National Institutes of Health or the University of Michigan.

References

8. https://github.com/biobanking/biobanking
10. https://github.com/EBISPOT/DUO
Evaluating a Visual Annotation Tool on Pediatric Weight Charts

P.J. Van Camp, MD1,2, Lei Liu1,3, C. Monifa Mahdi2, Andrew Spooner, MD2,3, Yizhao Ni, PhD2,3, Danny T.Y. Wu, PhD, MSi1,2

1Department of Biomedical Informatics, 2Department of Pediatrics, University of Cincinnati; 3Cincinnati Children’s Hospital Medical Center, Cincinnati, Ohio

Introduction

Weight is an important, vital sign in pediatric patients and often used for correctly dosing medications. Unfortunately, around 0.5% of the weights are entered incorrectly into electronic health record (EHR). These errors can lead to potential patient harm when under- or over-dosing medications based on these weights and thus capturing them as they enter the EHR would improve patient safety.

Statistical and mathematical methods already have been developed to evaluate pediatric weight charts and find outliers but we showed in previous work their performance is limited, especially in predicting new errors. Most weight-entry errors can be detected by medical experts by simply evaluating the weight charts, but time and workflow constraints make this impractical in clinic. By gathering a large collection of annotated charts, we could use supervised machine learning techniques to capture the expert’s intuition and build models with improved weight-entry error prediction. Building a large dataset of expertly annotated weight charts would be time-consuming without specific tools. Therefore, we developed a visual annotation tool (VAT) to aid in this process and present here the preliminary results of the first formal evaluation of this application.

Methods

The VAT utilizes visual analytics principles to improve annotation efficiency and accuracy. Specifically, it presents the participant with a grid of many charts at once, after which individual charts can be selected for detailed review and annotation. The annotation process has two styles: 1) one-step: annotate immediately after selecting a chart, or 2) two-step: first select all charts of interest and in a second pass annotate the selected charts one by one. Annotation is done by simply clicking weight points that are considered entry-errors. Coloring them red or orange will flag a weight as a definite or potential entry-error, respectively.

The user-centered evaluation was conducted in a lab setting. The participants, medical students who finished at least one pediatric rotation and pediatric residents, were assigned two datasets, each to be annotated with a different combination of settings for the size of the grid (3-by-3 or 5-by-5 weight charts), and annotation style (one-step vs. two-step). The two datasets contained 225 patient charts each and were randomly selected from a previously annotated dataset of 4,000 charts containing patients between 2 and 18 years old with at least 4 data points. This dataset had been annotated by three domain experts as reported in our previous work and would serve as a reference for error detection.

After testing, participants filled out the System Usability Scale (SUS), and 6 additional, tool-specific structured questions (scoring between 1 – 5 with 5 being highest) to provide feedback. Additional verbal feedback could be provided if clarification was needed. The experiment was reviewed and approved by the Institutional Review Board at the University of Cincinnati (IRB# 2017-2075) and supported by the Mary K. Logan research award received by Dr. Wu in 2018. A test version of the VAT can be accessed at http://3.16.217.8:3838/weed_dummy/

Results

This work reports the results of the first 10 participants (5 medical students and 5 pediatric residents). Seven of the participants were male. The average time to complete the first dataset (225 charts) was 9.1 minutes (SD 2.6) and 5.1 minutes (SD 1.6) for the second (significant time-difference, p = 0.001). The median number of points annotated over both datasets (550 charts) was 21 (range 16 - 39). Based on these results, it is estimated that an expert can annotate 1,000 weight charts in 30 minutes.

Table 1 summarizes the usability evaluation. The average SUS was 83.4 (SD 9.23). Scores above 68 are generally considered above average usability. For the structured questions (SQ), 5 out of 6 had an average score above 4. Based on SQ3 (score 3.1) our VAT could be improved in speed, though it is fast enough to complete the task without major impact on the usability according to the additional user’s feedback. Additional oral feedback also confirmed that the VAT supports rapid collection of expert annotation and was easy to use.
Table 1 – Results of the feedback after using the application

<table>
<thead>
<tr>
<th>Score Type</th>
<th>Avg. Score</th>
<th>Std.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUS composite score (&gt; 68: good usability)</td>
<td>83.4</td>
<td>9.23</td>
</tr>
<tr>
<td>SQ1. The use of a grid helped me to quickly find only charts that were of interest to me</td>
<td>4.7</td>
<td>0.46</td>
</tr>
<tr>
<td>SQ2. The display of the weight charts was intuitive considering my background on interpreting traditional weight charts</td>
<td>4.4</td>
<td>0.74</td>
</tr>
<tr>
<td>SQ3. The app allowed me all manipulations (viewing chart in detail, zooming, getting information on a single point, etc.) I needed for good analysis</td>
<td>4.5</td>
<td>0.52</td>
</tr>
<tr>
<td>SQ4. The color scheme was intuitive</td>
<td>4.4</td>
<td>0.74</td>
</tr>
<tr>
<td>SQ5. There was no significant lag or loading time that slowed down the workflow</td>
<td>3.1</td>
<td>0.88</td>
</tr>
<tr>
<td>SQ6. I encountered no technical errors</td>
<td>4.5</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Figure 1 shows the relationship between participants and the study experts regarding annotation of the 21 (possible) entry-errors defined by the latter. A ‘conservative agreement’ distinguishes between the two error labels (e.g. experts said ‘possible’ not the same as participant said ‘definite’), whereas ‘liberal agreement’ just looks if either error-label has been assigned. ‘Reference agreement’ compares how many participants agreed with the team’s experts, while ‘equal agreement’ looks at the most assigned label among all (participants and experts together, no comparing).

Figure 1. Agreement on annotated points (errors)

Discussion

An unexpected observation looking at figure 1 is that the wisdom of the crowd seems to be more powerful than an annotation by a select group of experts. For example, a 30kg increase in 3 days would be an obvious outlier and would likely be annotated as ‘definite error’ by anyone, expert or not. In contrast, a weight difference of 2.1kg in a five-day period could simply be missed by some annotators (both experts and participants) while others would judge that this could be medically explained and yet others would say it’s likely an error looking at the rest of the chart. In this case, the level of expertise in annotating is less important, given even experts could differ in opinion (as they did). By considering instead the number of participants out of the total that flagged this point, we can get some sort of confidence score (i.e. the more agree, the higher the confidence in the error) which can be very useful in later machine learning to assign a weight to each data-point instead of just the a label (e.g. A = error weight 100% (13/13), B = error weight 38% (5/13)).

In summary, we conducted a formal evaluation of a visual annotation tool (VAT) for weight chart annotation and showed it is easy to use and can support rapid collection of expert-annotated data. Analysis of the agreement between the participants and what was considered the reference highlighted the complexity of annotating weight errors and the necessity to utilize the wisdom of the crowd to identify potential weight errors considering their clinical importance. Once validated, our VAT can be used to annotate time-series data in other domains. Future work involves completing the analysis of evaluation data and revising the VAT to improve its efficiency and generalizability on other time series datasets.

References


1480
Patients’ Perceptions of Heart Failure Through the Lens of Standardized Nursing Terminologies

Haleh Vatani, MS, Sabita Acharya, PhD, Andrew D. Boyd, MD, Barbara Di Eugenio, PhD, Richard Cameron, PhD, Pamela Martyn-Nemeth, PhD, Debaleena Chattopadhyay, PhD, Pantea Habibi, BS, Carolyn Dickens, PhD, Amer Ardati, MD, & Karen Dunn Lopez, PhD
University of Illinois at Chicago, Chicago IL USA

Introduction

Effective communication between physicians, nurses, and patients is vital for patients to understand and engage in their self-care following hospitalizations. However, the use of different terminologies between physicians, nurses, and patients during hospitalizations can present a barrier to patients’ comprehension of health information. Our previous studies have shown only 21% of nursing concepts is used by physicians indicating lack of a common vocabulary between the two professions. Beyond the terminologies used by physicians and nurses when speaking to patients during hospitalizations, post hospitalization discharge summaries that inform patients of their self-care instructions often contain these terms that many patients find confusing. Although patients may be familiar with some of the physicians’ terms, research has demonstrated that patients make multiple types of errors when interpreting medical documents. Information on patients’ use and understanding of nursing terminology is limited. Although it is possible that patients are more familiar with terms used in nursing care documentation, there is a gap in knowledge of how often patients use terms that are synonymous with nursing care terms. These issues are particularly important for patients with heart failure (HF) because over 50% of HF patients do not have the required knowledge to understand their discharge instructions in order to take care of themselves at home. This lack of understanding is a contributing factor that makes HF one of the most common causes of hospital readmission within 30 days after discharge. The purpose of this study was to identify how often hospitalized patients with heart failure (HF) use terms that can be mapped to nursing terminologies, or their synonyms, when they describe their experiences with HF and health concerns.

Methods

As part of our larger research study (R01 CA225446-01), we conducted semi structured interviews with 21 patients (15 to 60 minutes, average 35 minutes) diagnosed with HF prior to their discharge from hospital. The research was approved by the IRB, participation was voluntary, and informed consent was obtained. The interviews, conducted by an expert in sociolinguistics, followed a narrative approach focusing on patient’s perceptions, feelings, and experiences of their disease. Interviews were audio-recorded and transcribed. Transcripts were preprocessed to remove sections where patients introduced themselves and answered general questions as well as the interviewers’ questions and responses. Then, the Clinical Text Analysis and Knowledge Extraction System (cTAKES) was used to process the remaining parts of the transcripts. cTAKES is a natural language processing system that turns health-related data extracted from unstructured text into structured data by mapping written words (e.g., “RA”) to medical concepts (e.g., “rheumatoid arthritis”) based on standardized medical ontologies. cTAKES mapped the health-related terms patients used with the Unified Medical Language System (UMLS) ontology to extract the Concept Unique Identifiers (CUIs). UMLS is a comprehensive list of clinical concepts and terms from various controlled vocabularies that provides a mapping structure among vocabularies and enables linking the terms from various terminology systems.

Second, we counted how many times a CUI from the UMLS 2018 files occurred in our patient interviews. From the UMLS 2018 mappings of Nursing Diagnosis: Definitions and Classification (NANDA-I), Nursing Interventions Classification (NIC), and Nursing Outcomes Classification (NOC) structured terminologies, nurse related CUIs were identified. If the same CUI belonged to more than one terminology (e.g., C0043144 is both NANDA-I and NOC), we credited both terminologies due to the overlap. To be most inclusive of patients’ use of terms, we included health-related terms (CUIs) that were negated (e.g. “no chest pain” or “no weight gain”).

Results

Our sample included 62% females (13) and 38% males (8) with an age range of 28 to 83 years old (AVG: 56; SD: 12.45) from three different ethnicity including 1) African American (86%, 18), Hispanic/Latino (10%, 2), and Pilipino (5%, 1). From the 21 interview transcripts, we identified a total of 2516 health-related terms (CUIs), 749 of
which mapped to the standardized 3 nursing terminologies. Theses 749 nursing terms were classified as 45% NANDA-I, 2% NIC, and 53% NOC. On average patients used 120 (SD: 61.82) health-related terms (CUIs) per interview with an average of 36 terms that mapped to nursing terminologies or their synonyms (SD: 19.31). Their use ranged from 19% to 67% per patient (AVG: 31.56%; SD: 0.11).

Figure 1. Percentage of health-related terms used by patients.  
Figure 2. Percentage of nursing terms used by patients.

Discussion
We found that nearly one-third of the health-related terms used by patients mapped to nursing terms or their synonyms for diagnosis, interventions, and outcomes. The vast majority of the terms were either diagnosis or outcomes with very little use of nursing interventions terms. This is not surprising given that patients may not be focused or aware of the many interventions that nurses perform during their hospitalization. However, patients’ understanding of the interventions performed by nurses (e.g. fluid management) in the hospital that will be shifted to the patient post hospitalization is vital to effective self-care. The high frequency of nursing terms or their synonyms in patient narratives and lack of nursing care documentation in Electronic Health Records (EHR) suggest an urgent need for incorporating interoperable nursing care information in patients’ longitudinal record to avoid any potential for gaps in care or communication. Our future work will examine patient use of physician terms or their synonyms in patient narratives. However, if nursing terms are commonly used, using these terms to frame patient discharge instructions may be an effective way to improve the effectiveness of discharge instructions.

Conclusion
Patients are using nursing terms or their synonyms to describe their health. As national efforts to engage patients in their EHR, these terms must be in the patients’ longitudinal record to increase patients’ comprehension about their health and empower their involvement in self-care.

References
Patient Stratification Process for enabling Clinical Interventions

Kavishwar B. Wagholikar MBBS PhD 1,3, Samuel Aronson ALM, MA 3, Benjamin M. Scirica MD 1,3, Akshay S. Desai MD MPH 1,3,4, Shawn N. Murphy MD PhD 1,3

1Harvard Medical School, Boston, MA; 2Massachusetts General Hospital, Boston, MA; 3Partners Healthcare Boston, MA; 4Brigham and Women’s Hospital, Boston, MA;

Introduction
An essential prelude to precision medicine is the stratification of patients into subpopulations for increased diagnostic and treatment efficacy. By efficiently matching patients to the right care management programs, improved outcomes can be obtained at lesser costs and in shorter time. However, electronic health records (EHR) functionality to assist clinical decision-making is restricted due to bias in health records and the difficulty in synthesizing meaningful concepts from the records. To address these challenges, we have developed and implemented a process at Partners healthcare to stratify patients for initiating clinical workflows, as described in this presentation.

Methods
We implemented a data-driven computer-based process to facilitate initiation of clinical interventions in the domain of cardiovascular medicine. 1 The process was iteratively implemented in eight cycles over a one-year period. In each run of the process, a list of patients that are eligible for the intervention, was generated by querying EHR data. This list was manually screened for validated by ancillary clinical staff by performing an in-depth-chart review to categorize patients as truly eligible and ineligible. They initiated therapy for the validated eligible patients and provided feedback regarding ineligible patients. The objective of the overall process was to identify the entire segment/strata of patients in the hospital population that are eligible for the clinical intervention, while minimizing the manual effort of the clinical staff to perform chart review. The process was iteratively developed by addressing issues identified in each run, and the final version of the process is described below.

Results
The patient stratification process consists roughly of the following steps: i) create a data store limited to variables required for computing the clinical concepts, ii) query the datastore to identify segments of the patient population that most eligible for the intervention, iii) manual validation to identify errors in the query results and improve the algorithm to derive variables, and iv) update the datastore to execute the improved algorithms. Figure 1 shows the detailed steps, as described below.

Create concept ontology: The concepts needed for clinical decision making are listed in an ontology as shown in Table 1, along with the data-points in the EHR needed to compute the
variables. For instance, the criteria ‘low ejection fraction’, is a concept required for initiating therapy, which depends on extraction of ejection fraction from cardiology notes.

**Single Data store:** The EHR data for computing the ontology is imported into a single datastore. For instance the patients who have been previously screened are stored external to the EHR, similar to notes and there is often a dedicated death index database. Assimilating all the data into a single store is essential for querying the dataset.

**Derived variables:** All the clinical concepts needed for decision making at the point-of-care are not readily available in the EHR data, and they are derived from the raw data. For instance, body mass index (BMI) is computed from weight and height using the BMI formula. Similarly, we found that echocardiograms can be segregated from cardiology notes by searching for 37 of possible 629 metadata codes. The concept of ‘is the Ejection fraction low’ is derived by checking if the ejection fraction is less than 40%. The ejection fraction is itself determined by running a complex regular expression parser on an echocardiogram note, which depends on accurately detecting echocardiograms from the cardiology notes. Consequently, a large number of derived variables are created from the raw EHR data, by executing logic of varying complexity.

**Query data store to generate sub-cohorts and a consort diagram:** After assimilating all the necessary information in a single store, the queries for the eligibility criteria are now executed to generate i) sub-cohorts for manual screening/validation and ii) a consort diagram that provides a quantitative view of how the criteria segment the patient population. The latter enables the team to estimate the number of patients for enrollment. Figure 2 depicts a consort diagram that provide a high-level view of the population segments. Next, the cohort generated for screening is sorted by sorting patients in order of decreasing confidence. For instance, patients visiting the hospital in recent months are more likely to continue care at the hospital and will consent for the health intervention. Hence the last visit date can be used for sorting the patient list. Similarly, a probability score output by ML can also be used for sorting.

**Learning from manual screening results:** The clinical staff now perform an in-depth chart review for the patient strata (segment) identified to be eligible for the intervention, tagging the reasons for ineligibility, along with the particular variable that was ‘derived’ inaccurately. Analysis of the screen-out reasons tagged by the clinical staff, helps identify inaccuracies in the data and the derivation. Moreover, additional data elements are identified that can improve the detection of eligible patients. Accordingly, the concept ontology is modified, and new EHR data is pulled into data-store, and the derived variables are recomputed.

**Discussion and Future work**

Bias and noise in health records poses a challenge to derive meaningful concepts from electronic health records, often making manual screening necessary to accurately identify the patient strata that is eligible for intervention. From an intensive validation approach, we decomposed the stratification process into data processing constructs, which can be individually fine-tuned to accurately identify eligible patients. Using an iterative approach that couples data-driven patient stratification with manual validation, helped scale the clinical intervention to a wider population, while minimizing the effort for the manual chart review. We continue to refine the stratification process and are currently implementing tooling to aid the process. Use of advance phenotyping techniques including unsupervised learning and active learning can further augment the efficiency of stratification.

**References**

**Acknowledgements**
This study was supported by NIH grants R01-HG009174 and R00-LM011575, and funding from Partner Healthcare
Dynamic Reaction Picklists for Improving Allergy Reaction Documentation

Li Qin Wang, PhD1,2, Suzanne V. Blackley, MA3, Carlos Ortega, BS1, Diane L. Seger1,3, Zfania Tom Korach, MD1,2, Kenneth H. Lai, MS4, Foster Goss, DO, MMSc5,6, Paige G Wickner, MD, MPH1,2, Kimberly G. Blumenthal, MD, MSc2,7, Li Zhou, MD, PhD1,2 1Brigham and Women’s Hospital, Boston, MA, USA; 2Harvard Medical School, Boston, MA, USA; 3Partners HealthCare, Boston, MA, USA; 4Brandeis University, Waltham, MA, USA; 5University of Colorado School of Medicine, CO, USA; 6University of Colorado Hospital, CO, USA; 7Massachusetts General Hospital, Boston, MA, USA

Introduction

Many allergies and adverse reactions need to be recorded into the electronic health record (EHR) allergy section to inform future prescribing. It is critical to obtain a complete and reliable allergy history for each patient and provide clinicians with efficient medication clinical decision support, e.g., drug-allergy alerts. However, the allergy modules in most existing EHRs have limitations in allergy and adverse reactions documentation and subsequent alerting. Although EHRs allow structured/coded entries, most current EHR allergy modules rely on commercial or local dictionaries whose adverse reaction lists often contain a limited number of reactions using a single, static pick list ordered alphabetically. Lack of a reaction terminology expressive enough to encode diverse reactions encountered in practice may lead to incomplete or free-text documentation. Recently, half of the reaction fields in the EHR were determined to be blank, and one sixth of reactions were entered as free-text. Incomplete allergy lists inhibit the triggering of necessary alerts, while inaccurate allergy entries will trigger unnecessary or inconsequential alerts. For example, any free-text entry will trigger allergy alerts with high-severity reaction regardless of its content, contributing to alert fatigue. In the present study, we investigated a dynamic reaction pick list using an enhanced reaction lexicon. With a dynamic pick list, once an allergen is entered, the reaction pick list will be populated with the most likely reactions based on that allergen. We hypothesize that the dynamic reaction pick list will allow providers to more efficiently select the correct reaction, thereby reducing the number of free-text entries.

Methods

Reaction List Development. In a previous study, we developed a comprehensive reaction value set of 787 reaction concepts using both coded and free-text entries over three decades from Partners Healthcare’s (PHS) Enterprise-wide Allergy Repository (PEAR), enhanced with adverse drug reactions data from two other institutions. The customized reaction picklist in our current EHR (Epic®) contains 49 reactions, including commonly reported reactions (e.g., rash, hives, itching) and serious reactions (e.g., anaphylaxis, Stevens-Johnsons syndrome). We merged and refined the reaction value set and the EHR picklist, resulting in 697 reactions to incorporate into the dynamic pick lists.

Allergy Entry Retrieval. We extracted all allergy entries from the inception of the PEAR database to October 8, 2018. Inactive entries and entries consist of “NO KNOWN ALLERGIES” were excluded.

Reaction Reconciliation. With our NLP tool, MTERMS, we extracted reactions from the free-text comment field in PEAR. For each allergy entry, we excluded any reactions found in free-text that were also among the coded reactions for that entry. This resulted in an updated table where each allergy entry contained all unique coded and free-text reactions associated with that allergen.

Allergen Consolidation. Our EHR contains 42,027 allergens, among which are many synonyms, brand names, or other variations. We used PHS’s mappings between allergens and medications and between medications and RxNorm ingredients to identify allergens with same substance(s) but different names, which we then merged into one allergen. All allergens at the drug class level (e.g., penicillins) were excluded from consolidation. We then generated an allergen and reaction co-occurrence table using the reconciled reactions and consolidated allergen list for developing dynamic reaction pick list.

Developing Dynamic Reaction Pick Lists. Based on the allergen-reaction entries, we developed dynamic reaction pick lists. We explored three approaches to rank reaction concepts with respect to a given allergen substance: support, lift (interest), and derived term frequency inverse document frequency (i.e., tfidf'). Figure 1 shows the formulations of the three measures we used to develop the dynamic reaction pick list.

\[
\begin{align*}
(1) \quad \text{support}(x, y) &= P(x,y) \\
(2) \quad \text{lift}(x,y) &= \frac{P(x,y)}{P(x)P(y)} \\
(3) \quad \text{tfidf}'(x, y) &= \frac{f_{x,y}}{1 + \sum_{i \neq x} f_{i,y}} \cdot \log \left( \frac{N}{1 + n_y} \right)
\end{align*}
\]

Figure 1. Formulations of three measures of strength of associations between allergens and reactions

1485
dynamic pick lists, where \( x \) is a reaction, \( y \) is an allergen, \( P(x) \), \( P(y) \), \( P(x,y) \) are the proportions of allergy entries with \( x \) reaction, \( y \) allergen and co-occurring \( x \) and \( y \), respectively, \( f_{x,y} \) is the frequency with which \( x \) and \( y \) co-occurred in the database, \( \sum_{x \in X} f_{x,y} \) is the total number of allergy entries with \( y \) allergen, \( N \) is the total number of allergens, and \( n_x \) is the number of allergens that co-occurred with reaction \( x \).

**Results**

We obtained 3,310,432 active allergy entries. After reaction reconciliation and allergen consolidation, those allergy entries included 2,500,055 allergy-reaction pairs. We generated dynamic pick lists for 9,992 allergens included in at least one allergy entry. Table 1 provides two example drug allergens, amoxicillin and ibuprofen, with their top 10 reactions ranked by support, lift, and tf-idf'.

**Table 1. Top 10 most common reactions for amoxicillin and ibuprofen sorted using three types of measures.**

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Support</th>
<th>Lift</th>
<th>tf-idf'</th>
<th>Reaction</th>
<th>Support</th>
<th>Lift</th>
<th>tf-idf'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rash</td>
<td>Morbilliform eruption</td>
<td>Rash</td>
<td>Gl Upset</td>
<td>Support</td>
<td>Lift</td>
<td>tf-idf'</td>
<td></td>
</tr>
<tr>
<td>GI Upset</td>
<td>Rash W/ Joint Pain</td>
<td>GI Upset</td>
<td>Rash</td>
<td>Support</td>
<td>Lift</td>
<td>tf-idf'</td>
<td></td>
</tr>
<tr>
<td>Swelling</td>
<td>Sores in throat</td>
<td>Diarrhea</td>
<td>Swelling</td>
<td>Rectal hemorrhage</td>
<td>Rash</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N/V</td>
<td>Erythema Multiforme</td>
<td>N/V</td>
<td>N/V</td>
<td>Upper GI hemorrhage</td>
<td>Angioedema</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swelling</td>
<td>Maculopapular Rash</td>
<td>Itching</td>
<td>Anaphylaxis</td>
<td>DRF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diarrhea</td>
<td>Candidiasis</td>
<td>Swelling</td>
<td>Itching</td>
<td>Kidney disease</td>
<td>GI hemorrhage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaphylaxis</td>
<td>TR</td>
<td>Anaphylaxis</td>
<td>Angioedema</td>
<td>Aseptic meningitis</td>
<td>Anaphylaxis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOB</td>
<td>Vaginitis</td>
<td>Candidiasis</td>
<td>SOB</td>
<td>VWD</td>
<td>Bleeding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angioedema</td>
<td>Illness</td>
<td>Angioedema</td>
<td>GI hemorrhage</td>
<td>Hematemesis</td>
<td>SOB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N/V: Nausea and/or Vomiting; TR: Transfusion reaction due to serum protein reaction; SOB: Shortness of Breath; Rash W/ Joint Pain: Rash with Joint Pains (Serum Sickness); GI: Gastrointestinal; BSUS: Bleeding stress ulcer of stomach; DRF: Decreased renal function; VWD: von Willebrand disorder.

**Discussion**

The three approaches ranked reactions differently. Support tended to rank frequently occurring reactions more highly, while lift gave a higher rank to reactions specific to the allergen in question. The reactions picked by tf-idf' and support were overall similar, but tf-idf' tended to put a lower weight on common reactions. For example, in the dynamic pick list ranked by tf-idf', the reaction itching more frequently co-occurred with allergen amoxicillin compared to the reaction diarrhea but was ranked lower as it was associated with more of other allergens. Further investigation is required to evaluate the impact of dynamic reaction pick lists on allergy module free-text allergy entries, clinical workflow, and user satisfaction, which can further inform which ranking mechanism is optimal for clinical documentation of patient allergies.

**Acknowledgements**

This project was supported by grant number R01HS025375 from the Agency for Healthcare Research and Quality.

**References**

Introduction

Medication information encoded in clinical notes within EHR can be critical for research on drug exposure response relationship to develop strategies to improve patient treatment. Existing medication extraction algorithms have been more general purpose, intended for extracting all medications within a piece of clinical text. A more targeted system that has the flexibility to adapt to idiosyncrasies in prescribing and writing patterns could ultimately lead to higher quality research datasets for studies including pharmacokinetic or pharmacogenomic analyses. We describe a natural language processing (NLP) algorithm developed using R called “medExtractR” to extract medication information such as strength, dose amount, and frequency from clinical notes. R is commonly used by many researchers, particularly statisticians, and at the time of submission there was no medication extraction package available. The medExtractR software is available for download as an R package from the Comprehensive R Archive Network (CRAN). We also compared medExtractR to three existing NLP systems for medication extraction: MedXN version 1.0.1, MedEx version 1.3.7, and CLAMP (Clinical Language Annotation, Modeling, and Processing) version 1.4.0.

Methods

MedExtractR was implemented using the R programming language. It relies on a combination of lexicons, rules, and regular expression patterns to identify relevant medication information. Lexicons for drug entities such as frequency were manually curated from expressions observed in training notes. The regular expressions were initially created by analyzing the contextual format of drug regimen information in manually annotated clinical notes in the training set. These were then iteratively refined to produce maximum performance on the training set. The system is designed to extract particular medications of interest, rather than all possible medications mentioned in a clinical note. Once medExtractR identifies a drug mention of interest within a note, it searches within a surrounding window for drug entities including drug name, strength, dose amount, and frequency. MedExtractR has function arguments which allow it to be customized for the drug of interest, including the size of the search window around a drug name and edit distance allowance for fuzzy matching of drug names.

We developed medExtractR on two medications, tacrolimus and lamotrigine, whose prescribing patterns ranged from simple to highly complex. We identified clinical notes for patients being treated with each drug from Vanderbilt University’s Synthetic Derivative, a de-identified database of clinical records derived from Vanderbilt’s EHR system. From these, we randomly selected sets of 60 training and 50 test notes for each drug. To create a gold standard dataset, we used manually annotated drug entities based on annotation guidelines developed using a double-annotation process. Inter-annotator agreement was assessed separately for tacrolimus and lamotrigine on a set of 20 clinical notes, each, using Cohen’s kappa. Performance was assessed using precision, recall, and F1-measure (F1). Additionally, we tested all four NLP systems on 110 clinical notes for an additional drug, allopurinol, to see how they worked on a novel drug not used in medExtractR development. In order to fairly compare medExtractR with the existing NLP systems MedXN, MedEx, and CLAMP, we manually reviewed raw output from both the existing NLP systems and medExtractR to ensure compatibility during evaluation, for example by converting semantically equivalent partial extractions to a standardized format. To evaluate uncertainty in our estimates, we computed 95% bootstrapped confidence intervals with 5,000 bootstrap replications at the note level.

Results

On 20 notes used for double annotation, the Cohen’s kappa for inter-annotator agreement for tacrolimus and lamotrigine was 0.970 and 0.837, respectively. Annotation guidelines were updated and clarified to resolve any instances of annotator disagreement prior to annotating the training and test sets. The training sets for tacrolimus and lamotrigine contained 60 notes each, with 105 and 102 drug mentions, respectively. MedExtractR achieved high precision/recall/F1 overall (macro-averaged over all drug entities combined) on the training set for both tacrolimus (.99/1.00/.99) and lamotrigine (1.00/.97/.98). The test sets for tacrolimus, lamotrigine, and allopurinol contained 50,
50, and 110 notes with 88, 76, and 191 drug mentions, respectively. Comparisons between medExtractR and MedXN, MedEx, and CLAMP were macro-averaged across each drug-entity (i.e., drug name, strength, dose amount, and frequency) pairing. MedExtractR often performed as well as or better than MedXN, MedEx, and CLAMP (Figure 1). For some entities (e.g., drug name for tacrolimus and allopurinol), medExtractR achieved lower precision than MedXN.

**Figure 1.** Entity-level comparison of medExtractR with existing medication extraction algorithms on test sets. Here, “n” refers to the number of annotations for that drug-entity combination in the gold standard dataset. Symbols and lines represent estimates and 95% bootstrapped confidence intervals, respectively. Arrows along the bottom x-axis indicate that either part or all of the confidence interval is below .80.

**Discussion**

Accurate and informative studies require high quality data, and thus the precise extraction of medication information from EHRs is a pivotal step in advancing research in the medical field. We developed medExtractR, a natural language processing system in R that performed better than three existing NLP systems (MedXN, MedEx, and CLAMP) for three different drugs, including one for which no training was performed. MedExtractR’s higher performance may be partially due to its ability to capture variations in writing patterns through customizable function arguments.

MedExtractR can be used to obtain more precise medication information that can provide quality data in diverse medication-based population studies, such as pharmacoepidemiology or pharmacogenomic studies. In order to improve medExtractR for use with other medications, we plan to implement expansions that will allow the system to extract drugs of different administration types, such as intravenous or oral solution medications. Future work will validate the system’s ability to correctly determine important clinical quantities including dose given intake and daily dose.

**References**

Using GPS to Locate the Missing and Track Wandering in People with Alzheimer’s Disease

Janusz Wojtusiak, PhD¹; Reyhaneh Mogharabnia, MS¹; U. Nalla B. Durai, MD¹;
Beverly Middle, DNP¹; Robert J. Koester³, PhD; Catherine Tompkins, PhD¹
¹George Mason University; ²Department of Veterans Affairs; ³dbS Productions LLC

Introduction

Alzheimer’s Disease (AD) and other forms of dementia constitute a large-scale public health concern with significant number of people with dementia being at risk of wandering and getting lost. These individuals may get hurt, cause distress to families and caregivers, and require costly search parties. Ubiquitous presence of IoT health technologies including health trackers, GPS devices, smartwatches and other wearables open new possibilities for improving safety and care for individuals with AD. In the presented work, we utilize GPS SmartSole by GTX Corp., one of several types of trackers available on the market, to assess feasibility of finding and predicting movement patterns¹ to aid in searching for missing people with AD and to better understand wandering. The GPS trackers work as long as they are charged and within cellular network range to send location data. This study addresses the issue of modeling movements so that a lost person can be found even if the device is not worn, not charged, or out of range.

The presented work is an extension of earlier study¹, in which we presented the possibility of predicting frequent locations for people with dementia. There are a number of other research projects related to the approach presented here, however, is unique because of its aims, scope and methodology. Some of the related works include research by Shoval et al.² to identify differences in movement patterns in people with dementia, Saltzman et al.³ that uses video-based observations to categorize wandering, further extended by Vuong et al.⁴ to implement a classification algorithm.

Our main objective is to study the application of machine learning methods to data from GPS trackers to create individualized models that describe patterns of movement, including frequent locations and local trajectories in the context of socioeconomic and AD specific information. Part of the process is construction of a decision support tool that helps to locate missing people with AD. The tool first predicts general possible area of search based on frequently visited locations. Then applies local trajectory modeling for specific patterns of movement in a given area.

Methods

Data Collection: Participants enrolled in the study are provided with GPS trackers. GPS data are extracted from the manufacturer’s information system. Caregivers are surveyed to establish baseline and resurveyed every 3 months, as well as asked to keep journals of wandering incidents. The survey questions include basic demographics, detailed information about incidents of being lost, information about caregivers, description of patient routines, detailed information about AD diagnosis, 13 questions specific to wandering, comorbidities in Charlson-Deyo categories, and medications used. Surveys are collected in Qualtrics and on paper, whichever form is preferred by caregivers.

Preprocessing: In the initial set of experiments sample data of 337 devices obtained from the GPS tracker manufacturer was analyzed. These pilot data were not linked to surveys. To date, additional data for 6 devices linked to surveys were analyzed. To protect the privacy of the individuals, we first de-identified data by randomly shifting and rotating all coordinates (such transformation still allows for spatiotemporal mining). We then converted timestamped geolocation data to weighted frequency domain as reported by Wojtusiak et al.¹. The data were then ordered by timestamp separately for each device. The first portion of the data was used for training and the 7 last days or 25% (whichever greater), was used for testing. This approach allows for simulating training and application of individualized models in realistic setting, as opposed to random data selection including cross-validation.

Spatiotemporal Clustering: We identified frequent locations using multipass version of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method, a clustering algorithm that is particularly applicable to geospatial data. In the first round, we identified top clusters with very high frequency in the data. We identified less frequent clusters in the following rounds. The multiple rounds of DBSCAN are needed because of large disproportion of weights between the most common locations and other locations, which also varies between devices. Our experiments indicate that the simple approach of 2-stage clustering procedure is typically sufficient.

Classification of Frequent Locations: Following the clustering analysis, we fitted classification models capable of predicting frequent location (given as a cluster assignment). In the first set of experiments we fitted models based on only two inputs: day of the week and time of the day. In the second set of experiments we included previous known locations coded as n-grams with n increasing from 1 to 5 and encoded duration of time spent in each cluster. Finally,
we simulated a delay between last known location and the time of prediction ranging from 1/2 to 5 hours, to model late search for a missing person.

We have tested a number of standard classification learning methods, including regularized logistic regression, random forest, naïve Bayes, deep neural networks and decision trees, as well as a baseline with Markov model. Hyperparameter tuning indicated that random forest with 100 trees achieved the best results. The model parameters were optimized in order to find tradeoff (Pareto solutions) between number of considered locations and model accuracy (AUC). Initial data preprocessing has been done in SQL and the all experimental work using Python 3 with Pandas and scikit-learn libraries.

Results

The applied spatiotemporal clustering led to stable results indicating frequent locations. The number of clusters varied from 1 to 9 between devices. Figure 1 illustrates identified frequent locations (marked in colors) for an example device. Points not belonging to any clusters (noise) are marked in black.

The obtained prediction results vary between devices. The performance of the method depends on specific individual being followed. In the first approach, for some individuals, based just on day of week and time the method is able to correctly predict 100% of outside home locations, indicating a very regular lifestyle. For others, the accuracy can be as low as 0%, indicating irregular lifestyle (frequent locations do not repeat over time).

The highest average prediction accuracy is achieved when using 2-grams previous known locations as inputs to the model. The average AUC is 0.774 +/- 0.214, accuracy 0.634 +/- 0.285, precision 0.669 +/-0.258, and recall 0.608 +/- 0.248 indicating moderate overall ability to predict locations. When considering delay in the search efforts, the AUC slowly drops to the value 0.724 when predicting with delay up to 5 hours. Slightly higher accuracies are obtained in devices with fewer frequent locations, but the number of clusters only partially explains variation in accuracy. Instead, low accuracies for certain devices are explained by very short time periods in which data were recorded.

Discussion

The presented results indicate the possibility of predicting locations of people with dementia using individualized models can contribute to search & rescue efforts when a person is missing. The presented research represents one step in larger project that aims at detecting and predicting wandering patterns. The models are being learned and applied in an adaptive decision support tool that learns movement patterns until certain accuracy is achieved. Our current efforts include using map data to probabilistically approximate likely trajectories from unequally spaced in time and distance GPS data reported by tracker devices.

Acknowledgement

The authors thank GTX Corp. for allowing access to raw data from GPS trackers needed for the presented study. This work has been supported in part by award no. 19-4 from the Commonwealth’s of Virginia Alzheimer’s and Related Diseases Research Award Fund, administered by the Virginia Center on Aging, School of Allied Health Professions, Virginia Commonwealth University.

References

Predicting Length of ICU Stay via Random Forest

Jingyi Wu, MS1,2,3, Ke Lin, MS1,2,3, Yu Lin, MS1,2,3, Yonghua Hu, PhD1,4, Guilan Kong, PhD2,3*
1Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China; 2National Institute of Health Data Science, Peking University, Beijing, China; 3Center for Data Science in Health and Medicine, Peking University, Beijing, China; 4Medical Informatics Center, Peking University, Beijing, China. *Corresponding email address: Guilan.kong@hsc.pku.edu.cn

BACKGROUND: The intensive care unit (ICU) provides advanced and resource-intensive treatment for the sickest hospitalized patients. The medical resources in ICU are usually in shortage compared to demands of ICU patients. Length of stay in ICU (LOS-ICU) is a key indicator for assessing efficiency of inpatient management. Prediction of LOS-ICU helps clinicians risk-stratify patients and allocate medical resources optimally. Although some studies have developed models to predict LOS-ICU, most of them are based on traditional methods such as logistic regression, and sample sizes are also limited in most studies.

METHODS: We used medical information mart for intensive care (MIMIC) III database1 for model development and validation. MIMIC-III contains medical records of 38,597 distinct adult patients admitted to critical care units between 2001 and 2012. In this study, we constructed three models: customized simplified acute physiology score (SAPS-II) model, classification and regression trees (CART) model, and random forest (RF) model for LOS-ICU prediction. The primary outcome was prolonged LOS-ICU (pLOS-ICU), defined as longer than the third quartile of patients’ LOS-ICU in the studied dataset2. Originally, the SAPS-II model is a tool used in ICU to assess the severity of illness for critically ill patients. There are a total of 17 variables used in the SAPS-II scoring system, consisting of age, type of admission, three types of chronic diseases, and 12 routine physiological measurements during the first 24 hours in ICU. Recently, some studies3,4 have customized the SAPS-II model for LOS-ICU prediction. The idea of building a SAPS-II based LOS-ICU prediction model is inspired by those studies. To compare performance of the three models objectively, all variables used in the customized SAPS-II model were used to construct the CART and RF models. Patients with variable missing rate larger than 30% were excluded. For those patients with more than one ICU admission during one hospitalization, we randomly selected one ICU admission record for each patient. Patients died within ICU were also excluded, as the LOS-ICU pattern of those patients who died within ICU may not be consistent with those patients who survived in ICU. The technique of Synthetic Minority Over-Sampling Technique (SMOTE) is a resampling approach to deal with imbalanced data. We used SMOTE to solve the problem of data imbalance in this study. We adopted 5-fold cross validation to evaluate prediction performance of the three models. The area under the receiver operation characteristic curve (AUROC), accuracy, sensitivity, specificity, Kappa value, positive predictive value (PPV), and geometric mean (Gm) were used as performance measures. Performance of the three models developed using raw dataset and the SMOTE processed dataset was compared. Performance differences between different models were assessed using two-sided t test.

RESULTS: After excluding 353 patients with variable missing rate >= 30% and 5160 patients died within ICU, a total of 39,162 ICU patients were included in the final cohort in this study. The observed average LOS-ICU was 3.8 +/- 5.7 days, with 23.6% proportion of pLOS-ICU. The proportion of male patients is 57.4%, and the average age of all patients is 61.8 +/- 16.6 years old. The prediction performance results, including the AUROC, accuracy, sensitivity, specificity, Kappa value, PPV, and Gm of the three models are compared in Table 1. Among the three models, the RF model achieved the most discriminative power with the highest AUROC (0.911), the highest accuracy (83.5%) and the highest Kappa value (0.670). The RF model also achieved the highest sensitivity (83.5%) and it outperformed the other two models in terms of the balanced accuracy Gm (83.3%), but the customized SAPS-II model achieved the highest specificity (98.1%). On the whole, the RF model had the best prediction performance taking into consideration of all performance measures. And we considered the CART model as the second best as CART yielded the second highest AUROC (0.808), the second highest accuracy (77.6%), and the second highest Kappa value (0.551). Table 2 shows the prediction performance of both the RF and CART models developed using the raw dataset and the SMOTE-processed dataset respectively. Models developed from SMOTE-processed dataset all performed better than those without SMOTE processing with statistical significance (P<0.001), especially in terms of balanced accuracy (Gm).

DISSCUSSION: Three models for LOS-ICU prediction were developed in this study. LOS-ICU prediction models based on the RF and CART outperformed the customized SAPS-II model, which is based on logistic regression algorithm. Using SMOTE method to deal with data imbalance made a substantial improvement of overall performance for the RF and CART models. SMOTE creates synthetic minority class examples based on the similarity on feature space, and can thus improve prediction performance of the derived models significantly. RF model showed the best...
discriminative power among the three models. Compared with the customized SAPS-II model, the RF and CART models were better at dealing with collinear data and therefore could achieve better predictive performance. Unlike CART model, RF model is an ensemble learning method. RF trains a multitude of decision trees independently and each tree is built using a bootstrap sample of the dataset. The output of the RF model depends on the voting result of all trees. Among the multiple trees, each one tries to model the relationship between the output and the covariates, and the RF model aggregates predictions of all trees to achieve a superior and more stable performance. The study results showed that RF based LOS-ICU prediction model had the best prediction performance among the three models based on the MIMIC-III dataset. Moreover, we confirmed the validity of using SMOTE to deal with data imbalance in prediction model development. However, this study has its limitations. First, the MIMIC-III database contains only patient data between 2001 and 2012, and has no data after. Second, selection bias may exist as we excluded those patients who died in ICU. Thus, the LOS-ICU prediction model developed in our study may not apply to those patients who will die in ICU. This study lays a foundation for future application of RF-based LOS-ICU prediction model in ICU clinical practice. It may help physicians make appropriate clinical interventions and do optimal medical resources allocation in ICU. However, further validation of its clinical effect is needed.

### Table 1 Performance comparison of three models

<table>
<thead>
<tr>
<th>Items</th>
<th>Customized SAPS-II</th>
<th>CART</th>
<th>RF</th>
<th>P*</th>
<th>P#</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC</td>
<td>0.668</td>
<td>0.808</td>
<td>0.911</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.766</td>
<td>0.776</td>
<td>0.835</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.072</td>
<td>0.770</td>
<td>0.778</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.981</td>
<td>0.781</td>
<td>0.892</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Kappa value</td>
<td>0.075</td>
<td>0.551</td>
<td>0.670</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PPV</td>
<td>0.258</td>
<td>0.715</td>
<td>0.794</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Gm</td>
<td>0.266</td>
<td>0.775</td>
<td>0.833</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

P* represents P value of t test between performance of CART and customized SAPS-II. P# represents P value of t test between RF and customized SAPS-II.

### Table 2 Performance comparison before and after SMOTE

<table>
<thead>
<tr>
<th>Items</th>
<th>before SMOTE</th>
<th>after SMOTE</th>
<th>P*</th>
<th>before SMOTE</th>
<th>after SMOTE</th>
<th>P*</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC</td>
<td>0.628</td>
<td>0.808</td>
<td>&lt;0.001</td>
<td>0.772</td>
<td>0.911</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.717</td>
<td>0.776</td>
<td>&lt;0.001</td>
<td>0.804</td>
<td>0.835</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.393</td>
<td>0.770</td>
<td>&lt;0.001</td>
<td>0.319</td>
<td>0.778</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.817</td>
<td>0.781</td>
<td>&lt;0.001</td>
<td>0.954</td>
<td>0.892</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Kappa value</td>
<td>0.211</td>
<td>0.551</td>
<td>&lt;0.001</td>
<td>0.333</td>
<td>0.670</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PPV</td>
<td>0.300</td>
<td>0.715</td>
<td>&lt;0.001</td>
<td>0.378</td>
<td>0.794</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Gm</td>
<td>0.567</td>
<td>0.775</td>
<td>&lt;0.001</td>
<td>0.552</td>
<td>0.833</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

P* represents P value of t test between results before and after SMOTE.

### References

Combining Publicly-Available and Electronic Health Record Data to Reposition Drugs

Patrick Wu, BS¹; QiPing Feng, PhD²; Joshua C. Denny, MD, MS¹,³; Wei-Qi Wei, MD, PhD¹
¹Department of Biomedical Informatics, ²Division of Clinical Pharmacology, ³Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232

Introduction
Repositioning of 'old' drugs to treat diseases is an attractive biomedical research topic, since it may substantially accelerate drug development by reusing compounds whose safety has been thoroughly examined in clinical trials.¹ We proposed to combine publicly-available results from genome-wide association studies (GWAS) and gene expression databases to find drugs that could potentially be repositioned to treat a new disease, followed by a validation analysis in the electronic health record (EHR). In this study, we demonstrated the feasibility of this approach to reposition drugs for hyperlipidemia. We found that lisinopril had an unexpectedly significant clinical effect on decreasing low-density lipoprotein (LDL) plasma levels. To our knowledge, this is one of the first studies to combine these unique publicly-available resources, contemporary computational tools, and EHR data for discovering drug repositioning candidates.

Methods
To impute hyperlipidemia gene expression signatures, we applied S-PrediXcan² on regression coefficients generated from a GWAS on LDL levels.³ We uploaded the top fifty (ranked by z-scores) up- and down-regulated genes to the integrated Library of Integrated Network-Based Cellular Signatures (iLINCS) platform.⁴ Pearson correlation coefficients were used to measure the similarity between the imputed transcriptome with the drug-induced expression patterns in the DrugMatrix library.⁵ We downloaded a list of drugs that generated only a reverse expression pattern from ≤ 1 incubation day.

This study’s objectives were to replicate known lipid-modifying drugs and validate a drug repositioning candidate in the EHR. To accomplish these aims, we removed the drugs not contained in the MEDication INDication high-precision subset (MEDI-HPS) database.⁶ Including only MEDI-HPS drugs allowed us to identify medications indicated for lipid disorders. From among the candidates, we manually curated drugs prescribed for treating hypertension, a common disease that provided us with a large starting patient cohort and adequate statistical power to detect differences in a drug’s effect on decreasing plasma LDL cholesterol levels.

To measure the target drug’s clinical effect on LDL levels, we used the Synthetic Derivative (SD), the de-identified EHR at Vanderbilt University Medical Center (VUMC). We identified adult patients who had ≥1 LDL measurement before and after they started the target drug (t₀), and had the drug listed on their EHR on ≥ 2 separate occasions. To prevent confounding by drugs known to alter significantly lipid levels, we removed patients under lipid-modifying therapy (e.g. statins, fibrates, and PCSK9 inhibitors). Since two drugs that we chose for validation were indicated for hypertension, we also removed patients taking other blood-pressure lowering drugs (e.g. calcium channel blockers, angiotensin II receptor blockers, ACE inhibitors, thiazide diuretics, and beta-blockers). For example, for the simvastatin group, we removed patients who had blood-pressure lowering drugs in their EHR and lipid-modifying medications (besides simvastatin) listed in their EHR. We then extracted the most recent LDL measurement before t₀ and all the LDL measurements when the patients were exposed to the drug, defined as being the period between t₀ and the last time that the drug was listed in their medical record.

We removed patients who had baseline LDL measurements <100 mg/dL. We then matched patients on age (± 3 years), LDL measurement before t₀ (± 5 mg/dL), sex, and race. For these patients we calculated change in LDL as the difference between their baseline LDL and the median LDL measurements after t₀. We employed permutation testing⁷ to compare the median LDL changes between the lisinopril cohort and two control groups. For the negative control comparison, our null hypothesis was that lisinopril cohort would have an LDL change that was equal to or less than the hydrochlorothiazide (HCTZ) group. For the positive control comparison, our null hypothesis was that the lisinopril cohort would have an LDL change that was more than the simvastatin cohort.

Results
Of the 308 drug repositioning hits, 71 (23.05%) had only reverse expression profiles with ≤ 1 day of incubation and
were contained in the MEDI-HPS database. From the 71, we identified four drugs that were FDA-approved for treating hyperlipidemia: fenofibrate, atorvastatin, lovastatin, and simvastatin. For the study’s second objective, we identified blood-pressure lowering compounds (captopril and quinapril) as potential candidates to validate in the EHR, as hypertension is a common disease and would thus provide a large patient cohort that would allow us to conduct a validation study using EHR data.

We chose lisinopril to validate this study’s approach in the EHR. Since lisinopril was not tested in the DrugMatrix database, we did not expect to see it appear as a repositioning candidate; however, it was the most commonly prescribed ‘ACE inhibitor’, to which captopril and quinapril also belong, in the SD. We chose HCTZ and simvastatin as negative and positive controls, respectively. Like lisinopril, HCTZ is a first-line blood-pressure lowering medication and was chosen as a comparator because its expression signature did not have a significant association with the imputed transcriptome. On the contrary, we selected simvastatin as the positive control comparator, as it lowers LDL and had a reverse expression pattern in DrugMatrix. In our comparison of the clinical effects of the three drugs on lowering plasma LDL cholesterol, we observed a statistically significant difference in LDL changes between the lisinopril with HCTZ (median LDL change = -5.5 mg/dL vs. 0 mg/dL; p < 0.05) and simvastatin cohorts (median LDL change = -5.5 mg/dL vs. -30.0 mg/dL; p<0.001) (Figure 1).

**Figure 1.** LDL changes for HCTZ, Lisinopril, and Simvastatin cohorts.

**Discussion**

In this study, we combined publicly-available databases, contemporary statistical tools, and EHR data to find and validate drug repositioning candidates for hyperlipidemia. A previous study used a similar approach to identify drug repositioning candidates for a number of brain-related disorders. We show that their pipeline can be replicated for non-psychiatric and neurological diseases and present a framework for validating these hits in the EHR.

**References**

Prevalence and Predictability of Low Yield Inpatient Laboratory Diagnostic Tests

Song Xu, PhD,1 Jason Hom, MD,2 Santhosh Balasubramanian,1 Lee Schroeder, MD, PhD,3 Nader Najafi, MD,4 Shivaal Roy,5 Jonathan H. Chen, MD, PhD1,2

1Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA; 2Division of Hospital Medicine, Department of Medicine, Stanford University, Stanford, CA; 3Department of Pathology, University of Michigan, School of Medicine, Ann Arbor, MI; 4Department of Medicine, University of California at San Francisco, San Francisco, CA; 5Department of Computer Science, Stanford University, Stanford, CA

Introduction

Lab testing is an important target for high value care initiatives, constituting the highest volume medical activity. It was estimated that up to half of all inpatient labs may be medically unnecessary1, but there is no systematic method to identify these in individual cases. Prior studies used machine learning to infer results of specific laboratory tests, such as ferritin2 and gamma-glutamyl transferase3, given results from other tests simultaneously available. Our initial investigation illustrated the predictability of machine learning methods on four laboratory tests at Stanford hospital4. In this work, we aim to identify low yield inpatient diagnostic lab testing based on more advanced machine learning methods that can make individualized estimates of likelihood on more than 90 high volume laboratory tests in a systematic and scalable way, validated on more recent data across datasets from multiple hospitals.

Methods

![Figure 1. Data processing, machine learning, and statistical analysis pipeline. Label “0” denotes the laboratory test actually generates a normal/negative result while “1” means the result is abnormal/positive. NPV: negative predictive value, PPV: positive predictive value, ROC: receiver operating characteristic.](image)

Data sources include structured electronic medical records from 110,000 inpatients at Stanford University Hospital in 2008-2017, 40,000 inpatients at University of Michigan (UMich) in 2015-2018, and 17,000 inpatient encounters at University of California at San Francisco (UCSF) in 2018. For each lab test, we randomly sampled 10,000 orders and separated them into training (development) and held-out test (validation) sets with a 75:25 split by patients IDs. We extracted over 800 features from both sets that are available at the time of the order entry, including patient demographics, most recent tests, Charlson comorbidity categories, common laboratory results, specialty teams treating the patient, and recent vital signs. Specialty team and vital sign data were not accessible in the UMich dataset, but predictions were still possible with other available contextual data. Counts of orders and comorbidities were binned into different time windows (last day, last week, last month, etc.). Numeric results of laboratory tests from last two weeks and vitals from last three days were summarized into various statistics (mean, max, most recent, etc.). We developed multivariable prediction models using various machine learning algorithms (Figure 1) and...
evaluated their performances in predicting the likelihood of a lab test order to yield a normal/negative result (defined by internal laboratory reference ranges). This likelihood score was then translated into a normal/abnormal final prediction by comparing against a decision threshold that was obtained by targeting at 95% negative predictive value in the training data. Statistical metrics including area under the receiver operating characteristic curve (AUROC = C-statistic), negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity were then calculated. We further evaluated the performance of models trained at one site but tested at another.

Results

The fraction of common inpatient laboratory tests that are repeated orders within 24 hours is plotted in Figure 2a. We evaluated the predictability of 76 high volume non-panel (standalone) labs and 15 common lab components (from “broad panels” such as Complete Blood Count and Metabolic Panel) at Stanford on whether they yielded a normal result. The best performing machine learning models generate >0.9 C-statistic for large volume standalone tests such as Troponin I, Lactate Dehydrogenase Total, and Sodium (Figure 2b) and common test components such as Hemoglobin, Platelet Count, Urea Nitrogen, Creatinine, Albumin, Total Protein, and Alkaline Phosphatase. For UCSF and UMich datasets, locally trained models outperformed models remotely trained at Stanford, but can still retain significant discriminatory power (C-statistic >0.85 for common test components).

Discussion

Low yield diagnostic testing is common and can be systematically identified through data-driven methods and patient-context-aware predictions. Our work is the first to do so by connecting advanced machine learning techniques to massive patient data, validated on a wide range of laboratory tests and across multiple hospitals. The method can be applied when highly predictable normal result is considered uninformative, including optimizing guideline and protocol-based testing and implementing precise best practice alerts in clinical decision support.

References

Computerized Conjoint Analysis of the Preferences of Patients with Serious Mental Illness

Alexander S. Young, MD MSHS1, Amy N. Cohen, PhD1, Nichole Goodsmith, MD1, Alison B. Hamilton, PhD1, Gerhard Hellemann, PhD1, Nancy Nowlin-Finch, MD1,2

1UCLA Department of Psychiatry and Biobehavioral Sciences, Los Angeles, CA, USA; 2Los Angeles County Department of Mental Health, Los Angeles, CA, USA

Abstract

Healthcare often involves choices among alternative treatments, considering patient preferences. It is not known whether automated assessment of preferences can be valid in patients with serious mental illness. We assessed preferences using a novel kiosk system tailored for patients with serious mental illness and conducted computerized conjoint analysis. This was implemented with 35 patients with schizophrenia at a public clinic. Computerized conjoint analysis produced valid assessments. The relationship between preferences and subsequent service use was complex.

Introduction

It is important that patients with common, serious mental illnesses such as schizophrenia have access to treatments that meet their preferences, and make informed choices among alternative treatments. Often, preferences are not elicited from patients and used to guide which treatments are implemented. Obesity is very common in serious mental illness and a cause of premature mortality. This population often takes medications that can cause substantial weight gain. People with serious mental illness die many years prematurely, mostly due to cardiovascular illness and cancer. In treatment guidelines for serious mental illnesses, there are multiple, different psychosocial treatments that result in reduced weight. None is widely used. If patient preferences were routinely assessed, clinicians and managers could know when to make alternative treatments available and for which patients. This project implemented and studied an innovative method for routine, automated assessment of treatment preferences. Marketing research has shown that choice-based methods of eliciting preferences, such as conjoint analysis, better predict revealed preferences, compared to direct questioning. However, people with serious mental illness usually have substantial cognitive deficits, can have limited literacy, and have somewhat less computer experience than the general population. The feasibility and validity of conjoint analysis in patients with serious mental illness (SMI) is unclear.1, 2

Methods

We developed computerized conjoint analysis for people with schizophrenia and implemented it at a public mental health clinic. Synthesizing the literature, interviews, and focus groups with 35 individuals with SMI, four key dimensions of weight services were identified: location (community or clinic); delivery mode (internet or in person); leader (clinician or layperson); and training mode (individual or group). A visual interface detailing choice options for these attributes was developed and presented to a focus group of 4 individuals with SMI. Enhancements were made and the design finalized. An additional 34 overweight, SMI-diagnosed patients at a community mental health clinic completed a computerized conjoint analysis session with the developed interface, and were administered the Computer System Usability Questionnaire (CSUQ) to assess user satisfaction and usability of computer systems (1-7 scale, lower indicates greater usability); the CSQ to assess satisfaction with services (1-4 scale, higher is more satisfied); and a qualitative interview about the conjoint session. They then were referred to a weekly weight management group at the clinic that was led by a clinician. After 3 months, patients were again administered the CSQ and completed a qualitative interview that explored their preferences for services as they related to the weight management groups.

CSUQ data were compared to established norms. Content analysis was performed on the qualitative data. We studied the validity of computerized conjoint preference assessment and whether the elicited treatment preferences predicted use of weight services. A multilevel logit model estimated both aggregate and individual level preference parameters from the partial preference data. The aggregate preference parameters show the overall weights that each specific feature of the intervention has in the logistic model predicting the choice between two interventions. The multilevel analysis framework also allows the estimation of the parameters of each individual participant’s logistic regression predicting their choice between any two interventions with known features. Using the individual participant’s preference models, we estimated their likelihood of choosing an intervention with the features of the provided in-
person weight management group compared to a neutral reference intervention with no features. Using this likelihood, we studied how closely the intervention that was provided matched their ideal intervention, and we report this as a treatment preference score.

Results
The average body mass index of participants was 35 (SD=5.1). Usability of the computer system was rated as excellent, with a mean overall usability rating of 2.01 (SD=0.99, scale mean: 2.82), system quality 2.04 (SD=1.04, scale mean: 2.8), information quality 2.12, (SD=1.17, scale mean: 3.02), and interface quality 1.9 (SD=1.11, scale mean: 2.49). Most participants praised the system for its ease of use, highlighting its visual aids, audio guidance, and touchscreen interface. However, many found its repetition of response options confusing or monotonous. Early feedback regarding the conjoint analysis system was used to simplify the system by reducing the number of characteristics per page, and to clarify administration by increasing indicators of the separation between pages. Seventeen participants utilized the weight management group, attending an average of 8.6 group sessions during the study period (SD = 8). Satisfaction with services on the CSQ averaged 3.5 (SD=0.6) at baseline, and 3.5 (SD=0.4) at follow-up. This is mid-way between good and excellent, and was unchanged over time.

Group-wide preferences had significant contributions from location (p<.001, favoring clinic), leader (p=.018, favoring clinician) and training (p<.001, favoring group), but not delivery mode (p=.68). Preferences did not correlate with age, gender, illness severity or body mass index. There were non-linear relationships among preferences and utilization of weight services (Figure 1).

Conclusion
Tailored computerized conjoint analysis can produce valid assessments of treatment preferences in patients with serious mental illness. Conjoint analysis could be a component of measurement-based care and inform decision-making. It can guide implementation of treatments consistent with patient preferences, and has the potential to increase use of effective treatments. The findings here may have applicability to other populations with cognitive deficits.

References
Classifying Non-Small Cell Lung Cancer Histopathology Types and Transcriptomic Subtypes using Convolutional Neural Networks

Kun-Hsing Yu, MD, PhD¹, Feiran Wang, MS², Gerald J. Berry, MD³, Christopher Ré, PhD⁴, Russ B. Altman, MD, PhD⁵, Michael Snyder, PhD⁶, Isaac S. Kohane, MD, PhD⁴
¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA; ²Department of Electrical Engineering; ³Department of Pathology; ⁴Department of Computer Science; ⁵Department of Bioengineering; ⁶Department of Genetics, Stanford University, Stanford, CA

Introduction

Non-small cell lung cancer is a leading cause of cancer death worldwide, and histopathological evaluation plays the primary role in its diagnosis. However, the inter-observer disagreement on this histopathological diagnosis is recognized, and the morphological patterns associated with the molecular subtypes have not been systematically studied. To bridge this gap, we developed a quantitative histopathology analytic system using convolutional neural networks to identify non-small cell lung cancer types and gene expression subtypes objectively.

Methods

We obtained the whole-slide histopathology images from 427 lung adenocarcinoma and 430 lung squamous cell carcinoma patients in The Cancer Genome Atlas (TCGA) database¹². Additional 87 patients with lung adenocarcinoma and 37 with lung squamous cell carcinoma were identified in the International Cancer Genome Consortium (ICGC) cohort. We retrained four convolutional neural network (CNN) models, including AlexNet¹³, GoogLeNet¹⁴, VGGNet-16¹⁵, and ResNet¹⁶, to identify tumorous regions from adjacent dense benign tissue and to classify tumor types. The models were fine-tuned by cross-validation on the training set. We divided the training set into two parts, trained the models with one part and examined the changes in performance in the other. Once we finalized the selection of hyperparameters, we trained a final model using the entire training set. We evaluated their performance by the areas under the receiver operating characteristic curves (AUCs) in the 20% held-out TCGA test set, validated the results in the independent ICGC cohort (n=124), and visualize the regions of importance in the classification process using the gradient-weighted class activation maps (Grad-CAMs)⁷.

To associate transcription-based subtypes of NSCLC with their histopathology patterns, level 3 publicly-available gene expression data of the tumor samples were acquired from the National Cancer Institute's Genomic Data Commons and the methods described by the TCGA Consortium¹² were used to determine the transcriptomics subtypes of each tumor. The associations between histopathology patterns and the transcriptomic subtypes were investigated by building a multi-class VGGNet-16 classification model. The output of the last layer of the VGGNet in the held-out test set was obtained and transformed using principal component analysis, with the first two principal components visualized. In order to evaluate the correlations between the variations in the transcriptomic signature that defined tumor subtypes and the histopathology-predicted subtype score, the Spearman’s correlation coefficient between the transcriptomic subtype score and the subtype probability predicted by histopathology images in the test set was calculated for each transcriptomic subtype. All statistical analyses were performed in R version 3.3.

Results

The CNN models successfully distinguished lung adenocarcinoma from the adjacent dense benign tissue (total patient number = 427), with the areas under receiver operating characteristic curves (AUC) approximately 0.941-0.965 in the TCGA test set. The results were validated in the independent ICGC cohort, with AUCs 0.890-0.935 (Figure 1A). For lung squamous cell carcinoma cohort (total patient number = 430), convolutional neural network classifiers achieved AUCs of 0.935-0.987 in distinguishing the tumor parts from the adjacent dense benign tissue in the TCGA test set. Similar performance was observed in the independent test cohort from ICGC (87 adenocarcinoma patients and 37 squamous cell carcinoma patients), with AUCs more than 0.979 (Figure 1B). Comparing with a previously-reported quantitative method for pathology analysis using the TCGA datasets⁸, our neural networks attained 6-12% performance improvement in the TCGA test set.

In addition, convolutional neural networks successfully distinguished adenocarcinoma from squamous cell carcinoma with AUCs in the TCGA test set 0.883-0.932, which were 12-27% better than the previously-described feature-based machine learning methods⁸. The AUCs of the finalized models applied to the independent test cohort (ICGC; total patient number = 124) were 0.752-0.857 (Figure 1C). The Grad-CAMs highlighted tumorous regions important for the tumor type classification (Figure 1D).
Discussion

We successfully classified the types and transcriptional subtypes of non-small cell lung cancer using convolutional neural networks. Compared with the previous methods\cite{6-8}, our approaches require no human segmentation or feature definition, making them easily fit into the clinical workflow of pathology diagnosis. In addition, we connected histopathology image patterns with gene expression profiles, which extended beyond previous works on single gene mutation\cite{3,4,5}. Furthermore, our models performed 6-27% better than the previously-proposed feature-based methods\cite{3}, and the results were validated in an independent cohort. The developed procedure is generalizable to other tumor types and can provide decision support to clinicians.

References

Identifying Factors Affecting Drug Discontinuation in Patients with Depression: Text Analysis of Patient Drug Review Posts

Maryam Zolnoori, PhD1,2,3, Che Ngufor, PhD1, Anthony Faiola, PhD4, Christina E. Eldredge, MD5, Jake Luo, PhD1, Sunghwan Sohn, PhD1, Joyce E. Balls-Berry5, PhD, Ahmad P. Tafti, PhD4, Nilay D. Shah, PhD2, Timothy B. Patrick, PhD5

1Dept. of Informatics and Administration, University of Wisconsin Milwaukee, Milwaukee, WI; 2National Library of Medicine, Bethesda, MD; 3Division of Digital Health Sciences, Mayo Clinic, Rochester, MN; 4Dept. of Biomedical and Health Information Sciences, University of Illinois at Chicago, Chicago, IL; 5School of Information, University of South Florida, Tampa, FL; 6Division of Epidemiology, Dept. of Health Sciences Research, Mayo Clinic Rochester MN; 7Dept. of Health Sciences Research, Mayo Clinic, Rochester, MN, United States; 8Dept. of Industrial and Manufacturing Engineering, University of Wisconsin Milwaukee, Milwaukee, WI

INTRODUCTION

The prevalence of antidepressant use among Americans increased from 7.7% in 1999–2002 to 12.7% in 2011-2014 (1), with the global market estimated at $11.6 billion in 2017 (2). The therapeutic benefits of antidepressants depend on adherence to the prescribed regimen, yet between 30 to 68% of patients discontinue their medications without completing an adequate course of treatment (3). This continuation can lead to poor health outcomes such as depression relapse, emergency department visits, low quality of life, and significant burdens on the individual, caregivers, family members and the health care system (4).

Although self-report scales (such as Adherence Barriers Questionnaire (ABQ)) for measuring patients’ discontinuation behavior towards antidepressants are well validated, they are associated with some methodological difficulties, such as not detecting potential factors and sampling bias. The premise of this study is that patient self-reports of their experiences with antidepressant therapy on drug review forums may constitute a reliable source to uncover various dimensions of antidepressants discontinuation that may not be identified by the self-report scales. To date, there are no reports in the literature that focus on identifying factors affecting antidepressants discontinuation in patients with depression as reported by patients in online healthcare forums. In sum, the objective of our study was to utilize patient self-report experiences, in online healthcare forums, to identify factors affecting patient discontinuation behavior of antidepressant medications in patients with depression.

METHODS

We used a novel mixed method approach to generate structured data from unstructured patients reports to test the hypotheses and model the relationship between antidepressants discontinuation and personal and healthcare variables. The detail of the methodology is as follows:

- **Identifying data source and drug source**: The data of this study collected from a healthcare forum called “askapatiens.com” that collects information on patients’ experiences with medications, age, gender, and duration of usage, and satisfaction with the drugs rated in a Likert scale from 1 (strongly dissatisfied) to 5 (strongly satisfied). For the purpose of this study, we considered patients’ satisfaction with antidepressants equivalent to patients’ attitudes towards antidepressants. A total of 892 patient reviews were randomly collected from the forum for the four most commonly prescribed antidepressants, including Sertraline (Zoloft) and Escitalopram (Lexapro) from the SSRI class, and Venlafaxine (Effexor) and duloxetine (Cymbalta) from the SNRI class.

- **Developing the Analytical Framework**: We used the Framework Method (7) with deductive-inductive approach to identify themes affecting patients’ discontinuation. In the deductive approach, we identified factors affecting patients’ antidepressants discontinuation toward antidepressants using thorough literature review. The factors were used as the initial themes for constructing the initial analytical framework for data analysis. In the inductive approach (open coding), 310 drug reviews were randomly selected for initial analysis using the initial analytical framework. Passages of drug reviews that could not be covered by the initial analytical framework were discussed in our regular team meeting for generating new themes. In the final step, themes were refined based on some rules, such as themes that covered less than 5% of the drug reviews were eliminated or merged with other themes. The identified themes were used for summarizing patients’ experiences with antidepressants and developing highly structured data from qualitative patient reports.

- **Analysis of dataset using the analytical framework**: In the first step of this phase, we addressed the grammatical and punctuation errors in the patient comments, and then split the comments into sentences using NLTK (a module from Python). All drug reviews in the sample were annotated using the analytical framework at sentence level. Four annotators with health backgrounds participated in the process of data annotation. All sentences were double-coded. **Calculating the Inter-annotator agreement (IAA)**: The Inter-annotator agreement (IAA) was calculated using Cohen Kappa. The overall IAA for the entire dataset was 0.75.

- **Strategy for handling missing values**: To handle the missing values, we eliminated all the drug reviews with no text. We also removed variables with low IAA and high number of missing values. To handle the rest of missing values, we adopted different imputation methods regarding the nature of the missing values for each variable. The details of the methodology for
developing this dataset were explained in our previous study with focus on identifying factors affecting patients’ attitudes towards medications (8).

RESULTS

Drug reviews were posted between February 2001 to February 2016. The majority of the patients (76%) were female. Approximately 30% of patients were less than 40 years of age. The majority of the reviews (63%) were posted by patients who had experience of using antidepressants for more than 3 months.

Testing the association between variables and drug discontinuation showed the following variables: “experience of adverse drug reactions” (ADRs), “perceived distress from ADRs”, “drug effectiveness”, “complaint about the lack of knowledge”, “patients’ attitude toward antidepressants”, and “duration of usage”, were all significantly associated with antidepressants discontinuation (P <0.05 for all listed variables). The variables of “age” and “gender” were not associated with drug discontinuation (P >0.05 for all listed variables). Moreover, modeling the relationship between variables and antidepressants discontinuation showed that variables “drug effectiveness” and “patients’ satisfaction with antidepressants” are the two most significant factors associated with drug discontinuation in patients with depression.

We also used a generated dataset to train a range of supervised machine learning classifiers to identifying text (annotated sentences) segments with comments related to adverse drug reaction, withdrawal symptoms, effectiveness, and ineffectiveness. Classifiers include: Support Vector Machine (SVMs), Random Forest (RF), Naïve Bayes (NB), and Logistic Regression (LR). The average F-score for SVM, RF, NB, and LR classifiers were 0.9, 0.87, 0.88, and 0.86 respectively, indicating that SVM classifiers outperform the RF, NB, and LR for this task.

DISCUSSION

The findings of this study have significant implications for developing clinical interventions aiming to improve patient adherence towards medications. One major implication of this study is that patients’ attitude is the most significant predictor of antidepressant discontinuation. This finding is in line with the literature showing that patient’s adherence behavior is strongly associated with patient’s attitude and belief towards medications. Clinicians are recommended to pay close attention to patients’ attitudes towards antidepressants and ensure that patients consider antidepressant usage as an essential part of their treatment. In addition, perceived drug effectiveness is a significant factor affecting drug discontinuation in patients with depression, and the full effect of antidepressant medications are not seen for typically four to six weeks. Therefore, clinicians are encouraged to track patients’ response towards antidepressants and encourage them to complete the adequate trials. Moreover, physicians should provide patients with adequate knowledge about antidepressants adverse effects and withdrawal symptoms. Providing patients with sufficient information on adverse effects can help them to have an informed decision about starting and continuing the medications and consequently reduce risk of abrupt discontinuation. In summary, the finding of this study may also help clinicians tailor interventions to improve their patients’ perceptions of their medications and consequently improve adherence to the antidepressant treatment.

REFERENCE

Early Prediction for Multiple Adverse Outcomes in Acute and Chronic Back Pain Transitions

Samir E. Abdelrahman¹, Michael G. Newman², Christina A. Porucznik²
Department of Biomedical Informatics¹ and Department of Family and Preventive Medicine ², School of Medicine, University of Utah

Introduction: Back pain is the most frequent cause of disability and inability to do activities of daily living. It usually results in dramatic social/economic/family costs, with an estimated U.S. prevalence of 27%¹. Prescribed medication may cause adverse outcomes including events (i.e., overdose, dependence, addiction) and side effects (e.g., sedation and respiratory depression). A back-pain patient may have multiple adverse outcomes (i.e., two or more outcomes that occur at the same time). Such outcomes are due to the interactions among the taken drugs that may introduce more pain complications or lead to a transition from acute to chronic pain. Clinicians still need to early predict the onsets of these multiple adverse outcomes to initiate a treatment or modify their treatment plan. Despite the various predictive analytics solutions in non-cancer or back pain domains, no study has been designed to satisfy this unmet need. Here, we propose to predict multiple adverse outcomes using the patient-specific data of the first day from encounter admission time for each back-pain transition.

Methods: We randomly selected 150,632 encounter (i.e., visit) data points from the University of Utah Healthcare (UUHC) electronic medical records (EMR) to train and test the predictive models. In this pilot study, we only used and analyzed a subset of patient features. For example, for back pain comorbidities, we only selected osteoarthritis, fibromyalgia, migraine, and headache as the most commonly-reported non-cancer chronic pains. For medications, we selected classes of opioids, nonnarcotic, and nonsteroidal anti-inflammatory agents as the most frequently used medications in the back pain literature. Furthermore, we only used repeated measures of the vital signs.

For each patient encounter, we included the following feature set. For the encounter outcome, we encoded the patient pain status into acute (A) or chronic (C) pain if the pain lasted shorter or longer than 3 months from the first patient encounter, respectively. Also, we encoded all occurrences events into E and all side effects into S such that each encounter had E/NE and S/NS to present event/no-event and side-effect/no-side-effect, respectively. Hence, our encounter outcome included 8 labels to present the 8 combinations of A/C, E/NE and S/NS, namely: A-E-S, A-NE-S, A-E-NS, A-NE-NS, C-E-S, C-NE-S, C-E-NS, and C-NE-NS. For the encounter predictor, we first used demographic information of age at admission time, patient class of inpatient/outpatient/emergency department, gender, ethnicity, visit types, and insurance classes. Second, we added 1-year prior information of the frequencies of back pain with 4 comorbidities, the frequencies of adverse outcomes, the counts of medication class orders, and the average of each vital sign repeated measures. Third, we collected the first 30-day repeated measures for each vital sign and discretized them into 30 features such that each feature had the average value of the repeated measures in the corresponding day. Finally, we imputed each day vital signs using the “last value carried forward” for the next patient days. If there is no patient value in that encounter, we used the literature clinical normal value of that vital sign to impute its missing value. To develop and test the trained models, we used 10-fold cross-validation and AutoWeka. AutoWeka is a widely used machine learning software package that validates different state-of-the-art classifiers with different hyper-parameter tuning to select the highest-performing classifier and the corresponding optimal classifier parameters. While running different AutoWeka classifiers, we also ran SubsetEval feature selection method and AntSearch algorithm to select the model with the highest area under curve (AUC) and the optimal number of features. We standardized all numeric predictors and used the one-hot encoding to encode each categorical predictor.

Results: Initially, we identified 86 predictors per day using our predictor encoding process. Among them, the selected predictors of the best predictive model performance were 32 including age at admission time, female, commercial insurance category, five visit types, all year prior information, and each day information of (prescribed medications, respiratory rate, temperature, and systolic blood pressure). The highest AUC resulted from random forest classifier of the first day with values 72.9% for A-NE-S, 79.1% for A-E-NS, 75.4% for A-NE-NS, 76.3% for C-NE-NS, 73.9% for C-E-NS, 70.5% for C-A-S, 62.3% for C-NE-S, and 75.5% for A-E-S with 1.5± 74.2 over the AUCs of all models’ at 30 days. We also conducted a comprehensive set of empirical experiments to assess the effect of predictors group combination by removing a combination of one or more group of predictors, i.e., demographics, prior, and current information. We found that removing priors decreased the models’ AUC by around 10% on average, and the worst results were obtained from the models that included only demographic information.

Discussion: Prior information has high predictive power. The data is imbalanced with respect to the 8 outcome values that led to moderate AUC results. Many patients were outpatients with length of stay ≤ 1 day; hence, the highest model performance was achieved using the first day data with priors. More advanced predictive analytics will be conducted.

References:
mHealth Tool for Patient-Centered Informed Consent

Fuad Abujarad, MS, PhD¹, Sandra Alfano, Pharm.D.¹, Peter Peduzzi PhD², Geoffrey Chupp MD¹

¹Yale School of Medicine, New Haven, CT, USA; ²Yale School of Public Health, New Haven, CT, USA

The objective of the informed consent (IC) process is to inform the patient or research subject about the purpose, procedures, risks, and benefits associated with clinical research and medical procedures.¹ However, the current IC process does not always guarantee full patient comprehension due to a number of barriers, including high-level medical vocabulary and lack of patient engagement.²-⁴ Therefore, we created an innovative IC communication process that overcomes barriers in IC communications and improves comprehension.

Our tool, Virtual Multimedia Interactive Informed Consent (VIC), was developed using a patient-centered design approach. We used front-end subject matter expert interviews and stakeholders focus groups with patients, researchers, and IRB members to develop a set of requirements. We completed a series of patient experience evaluations that mapped on to the standard development cycle, (i.e., user requirements analysis, design, implementation, testing, launch, and maintenance). To demonstrate VIC feasibility, we ran a two-arm randomized clinical trial to test VIC in an asthma study at the Yale Center for Asthma and Airway Disease Mechanisms.

We were able to surmount IC communications barriers with our innovative mHealth tool that runs on an iPad to communicate and deliver the IC process to patients in the asthma clinical trial. Information and messages were displayed on the iPad screen and spoken through headphones for patient privacy. An innovative feature of the tool is its ability to assess patient comprehension and offer a customized level of information. VIC educates and reinforces topics through the use of multimedia. The patient can view demos and presentations, listen to comments and explanations, get customized information, click on links to drill down for more information, ask questions and get answers, and rewind and replay as needed. We tested our tool using randomized control trial with 50 subjects in real world genetic clinical trial. Initial analyses of the study data suggest that comprehension was higher in the tablet informed consent group.

References
Efficient and Robust CT Image Segmentation with a Level Set Network

Jonas A. Actor, MA¹, Beatrice M. Riviere, PhD¹, David T. Fuentes, PhD²

¹Rice University, Houston, TX; ²MD Anderson Cancer Center, Houston, TX

Introduction

Medical image segmentation remains a difficult task: liver segmentation from abdominal CT scans is often done by hand, requiring too much time to use to treat hepatocellular carcinoma. Previously, image segmentation was done by solving the level set equation, a partial differential equation (PDE) describing how a boundary curve evolves given an image¹. Level set methods successfully capture clear edges and are robust to perturbation, due to the underlying structure of the PDE. In contrast, deep convolutional neural networks (DCNN) are susceptible to large changes in output given small perturbations of input². Despite this drawback, DCNNs are increasingly popular for liver segmentation.

Approach and Methods

We unrolled an iterative method for solving the level set equation, creating a level set network. In this framework, each iteration becomes a layer in a DCNN. The correct curve evolution is then ‘learned’ by the neural network. By definition, the output of this network solves the level set equation, and is therefore robust to perturbation. To test this method, we employed three segmentation methods on the MICCAI 2017 LiTS Challenge dataset³, consisting of 131 abdominal contrast-enhanced CT image stacks. These methods were: UNet⁴, a type of DCNN; ITK-SNAP⁵, a segmentation application using the level set equation; and our level set network (LSN). For UNet and LSN, we performed a 5-fold cross validation. In each fold, the selected LiTS data were split into training (90% of 4 remaining folds) and testing (10%). These networks trained via the Adadelta optimizer until saturation (40 epochs for UNet, 20 for LSN), with the Dice Similarity Coefficient (DSC) as the loss function. For ITK-SNAP, which allows users to hand-tune a few parameters, we allotted 10 min for a user to alter these parameters to achieve the best segmentation, after which ITK-SNAP solved the level set equation until DSC no longer improved.

Results

After performing the described methods, we obtained the following DSC scores.

<table>
<thead>
<tr>
<th>Method</th>
<th>UNet</th>
<th>ITK-SNAP</th>
<th>LSN</th>
</tr>
</thead>
<tbody>
<tr>
<td># Parameters</td>
<td>32M</td>
<td>3</td>
<td>69K</td>
</tr>
<tr>
<td>Avg DSC</td>
<td>0.955</td>
<td>0.745</td>
<td>0.804</td>
</tr>
</tbody>
</table>

Table 1: Parameters and DSC scores from three segmentation methods.

Conclusions

We observed that LSN achieved an average DSC score superior to ITK-SNAP, but below UNet. However, LSN guaranteed robustness to perturbation due to its analytical structure.

References

Unraveling Asthma Racial Disparities in Emergency Department Visit Using Electronic Healthcare Records and Machine Learning Algorithms

Adeboye A. Adejare Jr, M.S.,1,2 Yadu Gautam Ph. D.,2, Mekibib Altaye Ph. D.1,2, Tesfaye Mersha Ph. D.1,2
1University of Cincinnati, Cincinnati, OH, USA 2Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Asthma affects 12 million Americans, but the contributions towards asthma exacerbations do not equally affect all populations. African Americans (AA) are four times more likely to be hospitalized and five times more likely to die from asthma than European Americans. In addition, after being diagnosed with asthma, AAs have poor control leading towards more exacerbations. We developed a series of Machine Learning (ML) models including Logistic Regression (LR), Decision Tree (DT), eXtreme Gradient boosting (XGB) and Random Forest (RF), to determine asthma disparities between AA and EA patients in the Cincinnati Metropolitan area (Figure 1).

We leveraged data from University of Cincinnati (UC) electronic healthcare records (EHR) from the Emergency Department, pollution data from the Environmental Protection Agency (EPA) and environmental data from the Ohio Air Quality Agency (OAQA) between November 2012 and December 2017. We linked EHR clinical data together via unique deidentified identification numbers for a cohort of 42,374 patients (Table 1). We used the recorded entry date to merge the pollution, mold, and EHR data sets. This dataset was cleaned and fed to multiple machine models to predict both asthma severity and emergency visit disparities. FEV1% Values classified asthma severity, FEV1% ≥ 80, moderate (M), FEV1% 60-80, and severe (S) FEV1% < 60. Emergency Department (ED) visit frequency was counted for each person’s encounter in the ED in the time period. ED visitations below and at the third quartile for total visits by the cohort were classified as intermittent (I) visitors and visit frequencies above the third quartile as frequent (F) visitors. 

AAs samples were characterized by severe asthma and higher ED visits than their EA counterparts. AAs had a higher risk for environmental exposure with more elevated spores from mold and pollen than EAs. XGB produced the best predictive model for both AA and EA ED visit and severity (Figure 2). While pollution affected both groups, a higher predictive model was obtained for AAs than EAs. Environmental factors may contribute towards the disparities between races in asthma-related ED visits, and attention should be given to address these disparities.

### Table 1. EHR patient demographics

<table>
<thead>
<tr>
<th>Category</th>
<th>Overall</th>
<th>African American</th>
<th>European American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Population, Proportion</td>
<td>42,374</td>
<td>14,491 (34.2 %)</td>
<td>27,883 (65.8 %)</td>
</tr>
<tr>
<td>Male</td>
<td>13,903 (32.81 %)</td>
<td>4,955 (34.2%)</td>
<td>8,948 (32.1%)</td>
</tr>
<tr>
<td>Female</td>
<td>28,471 (67.19 %)</td>
<td>9,536 (64.8%)</td>
<td>18,935 (67.9%)</td>
</tr>
<tr>
<td>Age, Mean</td>
<td>49.6 (± 16.14)</td>
<td>48.16 (±15.47)</td>
<td>50.58 (±16.51)</td>
</tr>
<tr>
<td>FEV1%, Mean</td>
<td>72.19 (± 20.36)</td>
<td>69.18 (± 18.57)</td>
<td>73.90 (± 21.12)</td>
</tr>
<tr>
<td>ED Visit Frequency, Median</td>
<td>6 (1 - 229)</td>
<td>7 (1 - 215)</td>
<td>6 (1 - 229)</td>
</tr>
</tbody>
</table>

---

**Figure 1.** Development pipeline for data and machine learning algorithm implementation

**Figure 2.** Machine learning algorithm performance for ED visits and Asthma Severity by season. Evaluation was measured by accuracy of the model.
sureLDA: A Novel Multi-Disease Automated Phenotyping Method for the Electronic Health Record

Yuri Ahuja BS\textsuperscript{1,2}, Doudou Zhou PhD\textsuperscript{1,3}, Zeling He PhD\textsuperscript{1}, Jiehuan Sun PhD\textsuperscript{1,4}, Victor M. Castro MS\textsuperscript{5}, Vivian Gainer MS\textsuperscript{5}, Shawn N. Murphy MD PhD\textsuperscript{2,5}, Chuan Hong PhD\textsuperscript{1,2}, Tianxi Cai ScD\textsuperscript{1,2,4}

\textsuperscript{1}Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; \textsuperscript{2}Harvard Medical School, Boston, MA, USA; \textsuperscript{3}UC Davis, Davis, CA, USA; \textsuperscript{4}VA Boston Healthcare System, Boston, MA, USA; \textsuperscript{5}Partners Healthcare System, Charlestown, MA, USA

Introduction

Rich, multifaceted data from electronic health records (EHR) are increasingly available for clinical research. However, a major bottleneck in realizing the full translational potential of this data is its lack of reliable labels for many diseases. Though "silver-standard" proxies such as ICD billing codes are often used in place of true labels, these often suffer from dismal specificity - especially for rare diseases. On the other hand, traditional supervised phenotyping approaches require expert chart review to manually annotate labels, limiting scalability. Previous studies have introduced unsupervised machine learning algorithms using silver-standard labels. However, these methods only consider a single disease at a time and hence may not perform well when the goal is to phenotype multiple diseases simultaneously. We propose a fully unsupervised algorithm to efficiently predict multiple phenotypes for potential use in large-scale deep phenotyping.

Methods

Surrogate-guided ensemble LDA (sureLDA) method is an unsupervised multi-class prediction algorithm that first uses unsupervised clustering to obtain initial disease probabilities and then employs these probabilities to guide Latent Dirichlet Allocation (LDA) – a commonly employed topic model. Specifically, sureLDA sets the Dirichlet prior of the document-topic (i.e. patient-disease) distributions to the initial probability estimates, ensuring that learned topics align closely with the true disease labels. In this way, sureLDA successfully learns topics with important features specific to the disease phenotypes of interest. After training LDA and thereby obtaining disease-feature counts (i.e. disease scores) for each patient, sureLDA then combines these scores with surrogates via clustering ensemble to obtain robust final disease probabilities.

Results

In both simulations and EHR data from the Partners Healthcare Biobank, sureLDA consistently outperforms existing unsupervised phenotyping algorithms, particularly for rare diseases. Averaged over 10 diverse diseases in the Partners EHR dataset, sureLDA improves upon existing phenotyping algorithms MAP, PheNorm, classic LDA, and XPress with AUC increases of 0.023 (p=0.007), 0.029 (p=0.006), 0.071 (p<0.001), and 0.113 (p<0.001) respectively. It also outperforms supervised methods trained on 200-300 gold-standard labels.

Discussion

Robust to a variety of disease characteristics, sureLDA offers improved accuracy and precision over existing unsupervised phenotyping methods. It also exhibits powerful feature selection properties that make it relatively immune to high feature dimensions.

Conclusion

With high accuracy and multidimensional phenotyping capacity, sureLDA holds promise for large-scale phenotyping of the EHR, enabling more powerful utilization of the EHR for downstream study.
A Social Network Analysis Perspective of Comorbid Conditions with Chronic Kidney Disease

Ashwag. Alasmari13, MSc, Lina. Zhou2, PhD

1University of Maryland, Baltimore County, MD, United States; 2University of North Carolina at Charlotte, NC, United States; 3 King Khalid University, Abha, Saudi Arabia

Introduction

Comorbidity presents health management challenges to patients, healthcare professionals and organizations at large. Comorbid patients impose a high treatment burden in terms of understanding and self-managing their conditions, attending multiple appointments, and managing complex drug regimens1. Kidney disease is a common condition that occurs in patients with other chronic medical problems or complications2. Therefore, providing support for self-management of chronic kidney disease is crucial to patient care. Social media is increasingly influencing people’s daily life and their health behavior3. In particular, community Q&A sites provide a venue for generally asking questions and posting answers. However, researchers have yet to fully utilized Q&A sites to understand comorbid diseases and online patient community beyond what is already known in the healthcare setting. This study approaches the comorbid conditions that involve chronic kidney disease (CKD) through analyzing online Q&A data from a social network perspective.

Methods

We collected data from Quora, a popular community Q&A site. In addition to the questions, we also collected data based on the topics and subjects of the questions. We first compiled a list of topics related to the kidney disease using the snowball sampling strategy and then selected questions related to kidney disease, chronic kidney disease, kidney failure, dialysis, kidney-stones, polycystic kidney disease, CKD, dialysis patients, kidney infection, and pyelonephritis, etc. The dataset consisted of 1,312 questions in which were collected based on the selected topics. We grouped the topics into disease categories by mapping them to the general categories of ICD-10. The mapping results were further validated by a researcher with the medical background. Each node in the graph network represented a disease category and the size of the node indicates a higher weighted degree of the disease. The thickness of the edge indicates the strength of connection between two diseases that were mentioned along with CKD. We then used Gephi for analyzing and visualizing the disease network.

Findings

The network consisted of 42 comorbid diseases. The common comorbidities included in these findings were diabetes, hypertension, anemia, heart disease, depression, and thyroid disorder. The weighted degree were as following: diabetes=80, hypertension=73, anemia=31, heart disease=21, depression=11, and thyroid disorder=2. Hence, the two significant comorbid diseases were diabetes and hypertension. The findings complement previous research that involved medical records for kidney patients discharged from hospital with comorbid conditions3. Thus, our preliminary results demonstrate that social media serves as a useful source of information for patients. The social network analysis revealed a promising perspective for uncovering and understanding the underlying common diseases associated altogether. Additionally, the results complement the existing body of knowledge based on clinical findings. At the end, our research helps to give more realistic view on the complex nature of comorbidity associated with CKD.

References

Ocelot: A Modularized and Interactive Tool for Visual Understanding of Patient State

Gabriel A. Aldana, BBA1, Rohit Chaudry, MS1, Rimma Perotte, PhD1,2, Zachary M. Grinspan, MD3,4, Natalie Yip, MD5, David Vawdrey, PhD1,2

1Value Institute, NewYork-Presbyterian Hospital, New York, NY; 2Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY; 3Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, NY; 4Department of Pediatrics, Weill Cornell Medicine, New York, NY; 5Department of Medicine, Columbia University Irving Medical Center, New York, NY

Introduction

Following principles outlined by an international working group for bedside multimodal monitoring1, and working directly with the group’s chair, we designed a tool to display physiological and clinical data that we named Ocelot: Overview of Clinical Elements and Observations Timeline. Ocelot presents temporally-dependent heterogeneous physiologic variables together on one screen to help clinicians visually identify patterns, detect anomalies, and quickly draw inferences. The tool’s timeline view shows physiologic data in context with fluids status and medication administration data, including individual doses and continuous infusions.

Methods

The program is data source-agnostic and created entirely with open source tools such as the D3 JavaScript library, in order to promote portability. Ocelot was developed by a panel of physicians, informaticians, and engineers to meet the user demands for data in an intensive care unit. The tool development began in October 2017 and was placed into production in May 2018. There was no significant marketing campaign; users learned about the software through word-of-mouth. Ocelot is a real-time live tool (Figure 1) that constructs a visual display of vital signs, drips, medications, and ins & outs. The current version presents vital signs from nurse-validated flowsheet entries.

Results

In January 2019, Ocelot had 251 unique users, 60% of whom were physicians. Users viewed 425 patients and interacted with the application 1,019 times. Among the physicians, the most common user specialties were internal medicine (25%), pediatrics (20%), and surgery (10%).

Figure 1. A screenshot of the application. The user is able to hover over the line graphs to see individual recordings, filter out particular data elements, and collapse individual data types. A 24-hour summary of all vital signs is also displayed in a separate panel (not shown).

Conclusion

In future work, we will supplement the current set of displayed flowsheet metrics with real-time vital sign data drawn directly from bedside monitors. We will also iterate on the design of the application through qualitative and quantitative user feedback. Because we are logging application clicks and not just page views, we can identify which modules provide the most utility for different specialties, thus enabling the application to be personalized for the user. Given the modular design of the application, which can display data in a temporal and interactive format, we believe that Ocelot paves the way to being able to display many types of medical record data, and potentially even external data, to clinicians in a uniquely useful manner.

References

# SUPPORT-AF: Supporting Provider Prescribing of Oral Anticoagulant Therapy for AF with Anticoagulation Decision Support Using the Electronic Health Record

Azraa Amroze, BS\(^a\); Gordon Manning, MD\(^b\); Doug Buell, BA\(^b\); Alok Kapoor, MD, MSc\(^{a,b}\); David McManus, MD, MSc\(^{a,b}\); Jomol Mathew, PhD\(^b\)

\(^a\) Meyers Primary Care Institute, Worcester, MA, USA

\(^b\) University of Massachusetts Medical School, Worcester, MA, USA

## Introduction
Atrial fibrillation (AF) currently affects 5.2 million Americans, with 12 million projected by 2050. Anticoagulation (AC) therapy is the cornerstone for stroke prevention in patients with AF. However, only about half of eligible patients (i.e. CHA\(_2\)DS\(_2\)-VASc ≥ 2) receive AC in contemporary US-based, ambulatory registries. Researchers have assessed various interventions including decision aids, audit and feedback, and electronic messaging, but have achieved limited success in filling the AC treatment gap. A potential explanation for the lack of success is the absence of a tool integrated within the typical flow of practicing clinicians. We set out to leverage our EHR and design and evaluate a best practice advisory (BPA) tool positioned to trigger at the time of an ambulatory visit with a cardiology or primary care provider (PCP).

## Methods
We created a novel registry of AF patients using our enterprise research data repository – the UMMS Data Lake – linked with the Epic electronic health record (EHR) at our institution, the UMass Memorial health care system. To identify patients with AF eligible for AC, we used ICD-10 diagnostic codes, problem lists, medication lists, and labs available in the EHR. We further developed an algorithm that calculated each patient’s CHA\(_2\)DS\(_2\)-VASc stroke risk. In addition to being able to follow the patients, development of the registry enabled us to derive the exact logic to design a BPA using diagnostic and medication groupers that would fire whenever a patient with a CHA\(_2\)DS\(_2\)-VASc score ≥ 2 not currently on AC schedules an ambulatory visit with a cardiology provider or PCP. Once fired, we requested the provider to inform us of the reason for the patient not being on AC. If the provider did not declare a reason for withholding AC, he or she could select from a Smart Set order designed to guide prescription of an appropriate AC and is offered a link to a referral to our institution’s AC clinic. Included in the Smart Set was logic to offer options consistent with best practices; for example, direct oral AC would not be offered to a patient with renal failure. To validate the accuracy of the BPA, we also conducted chart-reviews on 100 randomly selected patients from our registry to ensure precision of AF status and stroke risk calculation. To evaluate the effectiveness of the BPA, we plan to assess the change in AC prescription percentage from the time of inception (baseline) to six months later for the group of patients for whom the BPA was triggered (treatment group) compared with those also seen at the medical center (by non-BPA providers such as surgeons or other medical subspecialists) for whom it did not fire. In order to control for secular bias, we will compare the difference in change across both groups and control for potential confounders such as CHA\(_2\)DS\(_2\)-VASc score and bleeding history/bleeding score.

## Results
We identified 6469 patients in our healthcare system with AF with a CHA\(_2\)DS\(_2\)-VASc stroke risk score ≥ 2, 1770 of whom were not on AC (27%) at the time of launch of our BPA in February 2019. We anticipate collecting change in AC prescription in August 2019. During chart validation, we found that our BPA was highly accurate in identifying AF (100%) and CHA\(_2\)DS\(_2\)-VASc score ≥ 2 (100%).

## Discussion
The EHR contains valuable information that can help address the practice gap in patients with AF eligible for AC. The BPA we designed was highly accurate in identifying AF patients with high stroke risk. In the near future we will be able to assess its effectiveness in increasing the rate of AC in eligible patients and also documenting the range of reasons why patients may not be on anticoagulation. Together these results will form the basis of a toolkit that can be adapted by other centers across the country.

![Figure 1. Data Lake for Shared Decision Support: Improve Prescription of AC among AF Patients at Risk for Stroke](image-url)
Using resilience engineering to understand an EHR transition

Shilo Anders, PhD, Patricia Sengstack, DNP, Carrie Reale, MSN, RN-BC, Laurie Novak, PhD, Joyce Harris, MA, Nancy Lorenzi, PhD, Elma Jashim, & Kim M. Unertl, PhD
Vanderbilt University Medical Center, Nashville, TN

Background. Large scale health systems that transition from one electronic health record (EHR) to another create a disruption to the organization, during which people must adapt and the environment must be redefined to accommodate the new EHR. This can also lead to numerous changes and even disruptions in clinical workflow, for which personnel can use resilient strategies in order to cope and adapt. Resilient performance is best understood as how people anticipate problems, monitor, adapt, and learn. From an engineering perspective, resilience is viewed as a way to offset when a system becomes brittle under production pressure or as a way to manage increasing complexity to minimize system failure or potentially catastrophic error. Organizations need optimal resilient performance from frontline personnel during an EHR transition. Thus, the goal of this research is to qualitatively examine the resilient performance strategies exhibited by a healthcare organization during and after a new EHR implementation.

Methods. Data were collected at 12 clinical sites (e.g. Adult Pulmonary Outpatient Clinic, Cardiac Inpatient Unit, Emergency Department) in 8 phases, beginning the day before implementation through 13 months post-implementation. Observation periods typically lasted 2 hours, although during Phase 1 of data collection (implementation day), observers were in place for up to 4 hours due to delays in the transition to the new EHR. Numerous observations were conducted at each location during each phase. During observation, the research team recorded free-text notes or had the option to record data on a workflow-focused data collection template. While observing, researchers conducted informal interviews with healthcare team members as time and circumstances allowed, typically asking questions related to the rationale behind observed behavior and on the impact of the new technology on work. Finally, observers collected artifacts related to the technology transition or to new processes related to the technology. All data were transcribed and uploaded into a web-based qualitative data analysis tool. Researchers coded documents for the use of resilient performance under the four “essentials” from the literature: 1) Learning – Knowing what has happened and training around system use; 2) Responding – Knowing what to do; 3) Monitoring – Knowing what to look for; and 4) Anticipating – Knowing what to expect.

Results. Researchers observed instances of all four categories of essential resilience strategies across all data collection sites. Along with consistencies between clinical sites, there were also variations in resilience strategies used immediately post-implementation versus after continued use of the new EHR that spanned all four themes. First, under the Learning category, new in-services were developed to help practitioners learn tips and tricks after rollout, even though the majority of the training had formally occurred prior to implementation of the new EHR. Informal knowledge learning occurred by asking peers with prior experience with the new EHR or with “super user” training. Secondly, in the Responding category, people throughout the organization experienced numerous issues and deviations across all observation phases and responded accordingly. Responses included: recruiting resources, resorting to and persistent use of paper, regressing to known prior EHR, and changing roles. For example, staff continually resorted to the previous EHR when they were short on time or had difficulty locating the information they desired in the new EHR. This was observed throughout the organization and persisted across all phases of observation. Thirdly, in the Monitoring category, researchers observed numerous people and processes designated specifically for monitoring emerging problems, as well as technologies that enabled individuals to report these problems. Lastly, in the Anticipating category, observations found varying responses at individual, unit, and organizational levels. For example, one nurse manager actually scheduled the most technology savvy nurses to work the day of the implementation on their unit.

Conclusion. Applying concepts of resilience engineering to the analysis of a large-scale health implementation affords a new framework to examine and understand the ways in which the organization and individuals adapt to handle disruptions to create a new routine. EHR system implementations continue across the nation. Understanding how resilience may impact success can provide organizations with important knowledge to prepare them for the challenge.

References
Managing Data Flows for Pediatric Complex-Care Patients and Their Families to Manage Care Plans

Steven B. Andrews PhD1, Mari-Ann Alexander PhD2, Justin Rousseau MD1, Anjum Khurshid MD PhD1, Jaimie Hancock1, Rahel X. Berhane MD2

1Dell Medical School at UT-Austin, Austin, TX 78712, USA
2Dell Children’s Medical Center, Austin, TX 78712, USA

Children with medical complexity (CMC) compose a sub-group of children with special health care needs (CSHCN) who face more hardships and have higher medical utilization. To support a mobile application providing shared care plans to enhance care coordination, our data team has been working to integrate data from a wide variety of sources. This poster describes the challenges and opportunities of bringing together these data to support care coordination and shared care plans for CMC.

Children with medical complexity (CMC) face hardships accessing medical care, often relying on tertiary hospitals for their care, and tend to be high utilisers of medical care in general.1-4 Care coordination has been suggested as a means of reducing medical utilization and helping CMC, but the results have been mixed. For example, while some studies have found significant results between care coordination and reduced ED visits, day-only admissions, reduced family travel time, and reduced school absences,5,6 others have found no improvement in health-related quality of life.6 Areas of research that have been emphasized are family-level interventions,1 integration of local health services with tertiary hospitals,7 and improving communication flows.7 Care planning interventions may help improve shared decision making for CMC, which has been shown to be important for improving quality care.2

Part of improving communications flows, increasing data and service coordination, and integrating the family more closely into the decision-making process is the use of a documented and easily accessed plan of care. Partnering with a commercial software development group, we have created a mobile app with the intention of providing care-related information to CMC and families, while also providing a mechanism for them to enter patient-reported outcomes (PRO) and manage a shared plan of care. This application is in production use, and we review some initial results.

This poster describes issues the team confronted bringing together data from multiple sources, including clinical settings with different EHRs (Cerner, Athena), payers, and PRO from the mobile app. Significant challenges were patient identification, patient matching, creating governance agreements to allow multi-directional data flows between partners, meeting and ensuring varying security requirements and processes, and working with the institutions on behalf of a population that comprises a small portion of the medical institution’s population. The poster summarizes lessons learned in unifying disparate data and providing a production environment to support families and clinicians.

References

5. Breen C, Altman L, Ging J, Deverell M, Woolfenden S, Zurynski Y. Significant reductions in tertiary hospital encounters and less travel for families after implementation of Paediatric Care Coordination in Australia. BMC Health Serv Res. 2018 Oct 3;18.
7. Cady RG, Belew JL. Parent Perspective on Care Coordination Services for Their Child with Medical Complexity. Children. 2017 Jun 6;4(6).
Differing patterns in frequency of electronic health records documentation among clinicians following the replacement of a legacy EHR system

Nate C. Apathy, BS1,2; Joshua R. Vest, MPH PhD1,2; Nir Menachemi, MPH PhD1,2; Justin G. Morea, DO3; Christopher A. Harle, PhD1,2

1Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, IN; 2Regenstrief Institute, Indianapolis, IN; 3Eskenazi Health, Indianapolis, IN

Introduction
Documentation is a substantial part of electronic health record (EHR) usage. Documentation includes clinical needs to record relevant patient information and provider activities, but regulations such as the Merit Based Incentive Payment System can increase the amount and frequency of documentation. Additionally, implementations may exacerbate this documentation burden, as users must learn how to use increasingly sophisticated EHR systems and adjust workflows while maintaining quality. Therefore, this study examined changes in users’ self-reported frequency of documentation in 4 settings before and after implementation of a new EHR system.

Study Design & Methods
We analyzed 4 items from a 2-year longitudinal survey of EHR users in a large public hospital system that switched from a homegrown legacy EHR to a vendor system in Oct. 2016. In each of 6 survey time points, users reported the frequency with which they entered data into the EHR across 4 scenarios: 1) clinical workdays – in patient rooms with the patient, 2) clinical workdays – between patients, 3) clinical workdays – evenings or nights, and 4) on days off. Each item had a 5-item Likert scale response set: Never, Rarely, Sometimes, Often, or All the time. We limited our study sample to clinical EHR users, divided into 3 groups based on their self-reported role: 1) Physicians, Physician Assistants, & Nurse Practitioners; 2) Nurses & Medical Assistants; and 3) other patient care roles. We calculated the percentage of users in each group reporting frequent use (Often or All of the time) for each item.

Results & Discussion
Our study sample consisted of 868 respondents at baseline, 821 at 1-month post-implementation, 1022 at 3 months, 787 at 6 months, 657 at 12 months, and 739 at 24 months. At baseline, 30% of Nurses and MAs reported frequently documenting in patient rooms with patients. After 1 month with the new EHR, this percentage increased to 72% and remained high. Similarly, at baseline, 22% of other patient care providers reported frequent use in patient rooms; this increased to 47% after 1 month and also held over 2 years. We found relatively little change in the percentage of physicians, PAs, and NPs reporting frequent use. Our findings suggest that EHR implementations may be related to increased documentation in patient rooms for non-physicians, although our analysis precludes causal interpretations. Our findings illustrate a varying burden of documentation between physicians and non-physicians, which can inform health system efforts to reduce the burden of frequent EHR use for specific user groups in different scenarios.

Figure 1. Percent of respondents answering “Often” or “All of the time” to time use survey questions, by clinical role
Generic vs. Specialized Models for Assessing Disabilities among People with Dementia
Negin Asadzadehzanjani, MS¹; Janusz Wojtusiak, PhD¹, Allison Williams, PhD², Cari Levy, MD³
¹George Mason University, Fairfax, VA; ²Department of Veterans Affairs, Bay Pines, FL; ³Department of Veterans Affairs, Denver, CO

Introduction
About 46.8 million of people are suffering from dementia and this number is expected to rise to about 130 million by 2050¹. Dementia impairs memory, thinking ability, comprehension and judgment and it can decrease functional abilities such as activity of daily living (ADLs)², thus accurate assessment of the functional ability helps decision making in patient’s care. Current assessment of ADLs is manual, costly and time-consuming; therefore, there is a need to develop an automated decision support tool to assess and predict ADLs. This work is an extension of our previous research on using EHR and administrative data to predict ADLs² and consists of two related components: comparison of dementia and non-dementia patients in terms of their disabilities and comparison of performance of machine learning-based models for automatically assess disabilities based on medical histories for people with dementia.

Methods
Demographic information, diagnoses and functional evaluations of patients over a ten-year period were extracted from their EHR and administrative data. Data was limited to patients with more than two evaluations. The diagnoses were mapped to CCS codes resulting in 281 distinct codes and coded as the number of days between the first (CCS_Max) and last day (CCS_Min) of the diagnosis and the time of evaluation (562 codes in total). The predicted outputs were nine out of ten categories of functional status (Bathing, grooming, toileting etc.) measured through Barthel Index³ and coded to indicate any level of disability. In the first stage, dementia and non-dementia people were compared for ADLs using propensity score matching. In the second stage, two types of models were developed to predict current ADLs according to patients’ diagnoses. The first (generic) model was developed using training data that included both dementia and non-dementia patients. The second (specialized) model was specifically trained on people with dementia. From a number of tested supervised learning methods Random Forrest was selected as the best algorithm. Hyperparameters were tuned through 10-fold cross validation. The two models were finally tested on people with dementia and compared using area under receiver-operator curve (AUC), accuracy, recall, and precision. Two-tailed t-test was used to determine if the performance of the models vary significantly (p<0.05).

Results
The cohort included 181,213 patients and 855,731 evaluations from which 79,466 (44%) were identified by 653 CCS code (Delirium, dementia, and amnestic and other cognitive disorders). The propensity score matching results indicated that people with dementia were significantly more disabled (0.66 vs. 0.55, p<0.01) than people without dementia on all ADL categories except walking (0.75 vs 0.78, p<0.01). High accuracy was achieved in automatically assessing ADLs using Random Forests with the average accuracy of about 90%. The average performance of the generic versus specialized model is as follows: accuracy 90.5% vs. 90.3%, AUC 94.7 vs. 94.6%, recall 92.3% vs. 92.4% and precision 92.8% vs. 92.4%. However, t-test revealed that the two models are not significantly different (p>0.05).

Conclusion
Our initial hypothesis was that specialized model would perform better than genetic model as the propensity score matching results indicated higher level of disability for dementia group. However, results revealed that the generic model performed slightly better than the specialized model even though the difference was not significant. The analysis of the learning curves indicated that the lack of difference was not caused by the size of the training data, but most likely by larger variance of data in the non-dementia group.

References
Which Patient Engagement Functionalities Are Salient for Quality of Patient Care?
O. Elijah Asagbra, PhD, MHA, CPHQ
East Carolina University, Greenville, NC, USA

Introduction
Hospitals are increasingly adopting health information technology (HIT) to facilitate patient engagement to meet governmental pressures,1, 2 as well as consumer demand for using technology in their health.4, 5 Several studies have highlighted the functionalities and benefits of HIT for patient engagement, and others have examined the association of HIT and care quality. However, few have determined if any one HIT functionality may be salient for improved quality of care. This study therefore aims to examine the relationship between the specific patient engagement functionalities offered through health information technology (HIT) and quality of care.

Methods
Data on hospital adoption of patient engagement functionalities obtained from the American Hospital Association (AHA) IT supplement were combined with quality data obtained from the Centers for Medicare and Medicaid Services (CMS) Hospital Compare. Hospital fixed effects regression models with time dummies were used to analyze a panel data consisting of 1,463 hospitals from 2012 to 2014. Quality of care was measured as patient satisfaction and 30-day readmission rate for acute myocardial infarction, heart failure, and pneumonia.

Results
This study revealed that certain patient engagement functions were significantly associated with improved quality of care measures in hospitals. For instance, it was determined that hospitals offering patients’ electronic access to patient–specific education materials (β= 0.806, p < 0.05), the ability to download their health information (β= 0.640, p < 0.05), the ability to request changes in their medical records (β= 0.778, p < 0.05), and the ability to generate their own health data (β= 0.832, p < 0.05) were associated with improved patient satisfaction. Additionally, hospitals allowing patients to view their health records and to download their health information were also significantly associated with improved readmission rates for acute myocardial infarction, heart failure, and pneumonia respectively. Permitting patients to access their discharge instructions electronically (β= -0.658, p < 0.05) was significantly associated with improved acute myocardial infarction readmission rates and allowing patients to generate their health data (β= -0.280, p < 0.05) was significantly associated with improved readmission rates for heart failure.

Conclusions
Hospitals can improve their overall quality by offering health information technology with certain engagement functionalities to their patients. While it is important to adopt a broad range of functionalities to meet internal and external demands, the findings of this particular study suggest that hospitals that offer patient engagement HIT functionalities related to patient education, viewing and downloading health information, requesting changes in medical record, and patient generated data entry are likely to have higher patient satisfaction and lower readmission rates for acute myocardial infarction, heart failure, and pneumonia. It may be because these functionalities enable patients to become more engaged and activated in their care. This study will potentially assist hospital administrators in justifying their strategic deployment of HIT resources to improve both perceived and actual care quality.

References
A checklist to support effective and scalable design of ePRO tools in electronic health records

Elizabeth Austin, MPH1, Courtney Segal, BA1, Joseph Heim, PhD1, Savitha Sangameswaran, MS1, Danielle Lavallee, PharmD, PhD1

1University of Washington, Seattle, WA

Introduction: Patient-reported outcomes (PROs) provide important information about how patients experience health conditions and treatments, as well as quality of life. When integrated into care delivery, PROs can provide clinically-meaningful data and ensure treatment decisions are guided by patient perspectives. Yet current use of PROs in clinical care is often limited to paper-based workflows that prohibit efficient, integrated reporting and data use. Therefore, many health systems are starting to utilize electronic PROs tools, often through electronic medical record systems (EMR). The electronic capture of PROs parallels many other types of patient-reported data currently in use, yet PROs also introduce nuanced challenges (i.e. adaptive logic, complex scoring and interpretation) that add complexity to the inputs, activities, and outputs designed in ePRO tools. While there is a growing evidence base on how PROs can inform care, evidence based practices for ePRO design and implementation are limited.1 As clinical and quality drivers continue to increase the demand for PROs, health systems need guidance on how to best govern the ePRO build and implementation process to support diverse clinical and user needs.1

Methods: Working with a multidisciplinary group (clinical, administrative, IT, research, systems engineering), we supported the design of four distinct PRO tools within our EMR (Table 1). The team identified that existing tools used to request IT resources did not adequately prepare project teams for ePRO builds. We used structured analysis and design technique (SADT, or IDEF0) methodology to identify core functions (i.e. activities) related to ePRO use. We then leveraged the four PRO tools as use cases to model related mechanisms and constraints for each core function. Learnings from the build and pilot experiences of each of the use cases were applied to ensure validity across clinical contexts and workflows. Lastly, we developed an ePRO implementation checklist to support the intake and design of future ePRO builds, based on workflow gaps identified via SADT modeling and pilot results.

Table 1: Clinical context of PRO tool pilots

<table>
<thead>
<tr>
<th>Use case</th>
<th>Clinical context and goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Joint Replacement</td>
<td>Pre and post-surgical assessment of functional status</td>
</tr>
<tr>
<td>Depression Screening</td>
<td>Population health screening for depression risk</td>
</tr>
<tr>
<td>Depression Management</td>
<td>Ongoing assessment of depression symptoms and response to treatment</td>
</tr>
<tr>
<td>Medicare Annual Wellness Visit</td>
<td>Multiple PRO and health status measures to identify preventive care needs</td>
</tr>
</tbody>
</table>

Results: SADT modeling helped identify five principal functions related to the use of ePROs in clinical care: deploying ePRO measures, collecting ePRO data, tracking ePRO completion, reviewing ePRO results, and storing ePRO data for future use. Evaluation of the four pilots emphasized the need to consider enhanced design choices beyond the patient-facing survey itself as part of the preliminary build intake assessment. In particular, activities related to the functional area of tracking were missed in all four build examples, and several critical activities related to all five functions were not addressed in existing IT build assessment tools. The ePRO implementation checklist translated modeling results into questions that would help teams better identify ePRO-specific dependencies (e.g., existing ordersets, optimization of local clinic roles) and constraints that might negatively impact the build and implementation process (e.g., EMR data lags). The checklist reviews elements of the ePRO project scope, data collection and submission workflow, and electronic tools needed to support ePRO data review and documentation.

Conclusion: As health systems seek to leverage their EMRs to engage with patients through the use of PROs, there is a need to understand best practices that can support effective and repeatable ePRO design and builds. SADT modeling across four ePRO use cases highlighted multiple considerations that impacted the success of ePRO tools. The resulting ePRO implementation checklist provides a standardized tool to guide teams in the assessment, prioritization, and governance of ePRO builds. Future work should identify useful metrics to monitor the effectiveness of ePRO tools, including understanding reach, workflow efficiency, and variation in clinical practice.

References

Solutions in Health Analytics for Rural Equity Across the Northwest

Uba Backonja, PhD MS RN1,2, Seung Eun Park, RN MSN2,
Anne M. Turner, MD MLIS MPH2, India Ornelas PhD MPH2, Greg Whitman, BA2,
Melinda Schultz, MS2, Betty Bekemeier, PhD MPH FAAN2

1University of Washington Tacoma, Tacoma, WA, USA;
2University of Washington, Seattle, WA USA

Introduction

Public health decision making requires access to data to identify health disparities, implement prevention efforts, and ensure health care access. Clinical care is being transformed by advancing analytic methods applying a large volume and variety of data. However, these benefits have yet to be realized in rural public health, which faces a lack of access to quality data to understand communities’ needs and is impacted by a lack of access to data that includes marginalized and underserved populations. These data gaps reduce rural public health’s capacity to identify and address inequities. Therefore, there is a need to support rural public health’s access to and use of data.

SHARE-NW: Solutions in Health Analytics for Rural Equity across the Northwest

The project: SHARE-NW is a 5-year, user-centered project funded by the U.S. Health and Human Services Office of Minority Health (tinyurl.com/sharenw). Our goal is to identify, gather, and visualize data to support rural public health agencies and leaders to more effectively understand and address inequities in rural Washington, Oregon, Idaho, and Alaska. The final product will be a website that supports access to data on health outcomes and social determinants of health (SDOH), provides visualizations of data to facilitate understanding of local health equity, and engages users in training opportunities regarding data use and equity. We are engaging end-users – rural public health agency leaders and data users – throughout the project lifecycle to support the development of a solution that meets their needs.

Completed work for SHARE-NW: In Year 1 (2017-18), we interviewed end-users and learned they lack capacity to access and use data, encounter data quality issues, and desire training on using data. We also compiled, and are continuing to compile, publicly available data on SDOH and health outcomes related to six priority areas for rural health: obesity, diabetes, tobacco, mental health (including substance abuse and suicide), violence/injury, and oral health. In Year 2 (2018-19), we developed paper mockups of visualizations using data related to opioids. Designs were informed by Year 1 findings, a review of online health data dashboards (e.g., countyhealthrankings.org), and the literature (e.g., Munzner’s visualization model). We evaluated mockups in-person, via phone, or videoconference with end-users. Initial analyses indicate participants found value in the dashboard to make data easier to access, use, and share. We also identified (a) indicators included in publicly available datasets matching our six priority areas and (b) publicly available trainings related to needs identified in Year 1. All project endpoints to date have been met.

Future work for SHARE-NW: In Year 3 (2019-20), we will build a visualization prototype in Tableau based on the work completed Years 1-2. This prototype will be tested among end-users and updated based on resulting feedback. We will also develop trainings in partnership with the Northwest Center for Public Health Practice (nwcpp.org). Trainings will include how to use the website and data for decision-making regarding health equity, as well as other topics identified through the project. In Years 4-5 (2020-22), we will continue usability testing, and update of the dashboard, provide trainings, identify more datasets to integrate, and develop strategies to sustain SHARE-NW.

Conclusion

There is a great need to develop informatics solutions to support rural health equity. SHARE-NW aims to address this need by engaging rural health practitioners to develop a solution to support data access and training. There is potential for this work to inform practices and policies regarding user-centered development of public health data dashboards.

References

Growing Up Stronger: Improving Pediatric Population Health with HealtheRegistries
Ranjodh Badh, Children’s National Medical Center, Washington, DC;

Abstract
Children’s National Health System set out a lofty goal aimed at improving the health of children in the region with five health conditions: Asthma, Cardiomyopathy, Diabetes, Inflammatory Bowel Disease and Sickle Cell Disease. Our workflow-integrated decision support tools guide our providers by alerting them to evidence-based improvement opportunities for children with these conditions. We describe how EHR-derived data aggregation, data quality improvement, and workflow integration can effectively improve population health outcomes. Our learning objectives are identifying data integrity challenges and best practices to overcome these with successful outcomes; understanding the relationship for improving data quality, workflow integration and improvement in regional pediatric health for chronic conditions.

Introduction
In 2017, CDC estimated 17,429 children in Washington D.C. have Asthma. Children’s National Health System provided services to 15,192 children, reflecting 87.5% data capture. In 2018, Children’s National Health System started with the goal to improve health of children in the region. As an example, we identified at-risk children with severe persistent asthma, with more than 20 ED Visits in past 3 years and multiple hospitalizations over the same period. Meaningful data from our population health analytic tools enabled coordinated interventions and communications for patients with certain (chronic) conditions, to empower patients and families to manage their disease and prevent complications. Success in population health is multifaceted and includes engagement of subspecialists caring for these children. Our analytics platform helps to identify opportunities for improvement, care coordination and management across the continuum, and clinical integration with the primary care team.

Two-arm Approach—Data Quality and Care Outcomes
Advancing quality outcomes in select population health diagnoses across the continuum of care through our patient registries drives our organization to focus attention on the measures of health on data quality and care coordination—our action arm. Data quality is challenging in a multifaceted manner:
- Health quality measure performed, but not in accessible or documented
- Care ordered but not carried out
- Measure documented in an unstructured (narrative) manner
- Mapping from EHR to enterprise data warehouse (EDW) is inaccurate and requires correction

The chart below displays progression of two measures over time.
1. The sudden increase/decrease in % of patients met is due to changes in data correction
2. Gradual improvement seen in both measures is weekly improvement made by clinical teams.

In our approach to each of the above issues we have seen some steady improvement in data quality over the course of the year. It is clear data quality is only part of the story. Our second arm, the clinical integration with Ambulatory Case Management, is a crucial role to improve outcomes of care align with support case managers in mobilizing care resources to mitigate risk of adverse events.

Conclusion
Transforming population health is central to our mission which requires reaching to region, inclusive of kids who may not seek out our care. Granting our clinicians provider facing analytic tool aims at helping our care teams take better care of our kids through data understanding.

References
Completeness and Concordance in Electronic Health Record Documentation of Chemotherapy-Induced Nausea and Vomiting in Pediatric Cancer

Melissa Beauchemin, MSN, RN,1 Rebecca Schnall, PhD, RN, MPH,1 Chunhua Weng, PhD, MS2

1School of Nursing, Columbia University, New York, NY
2Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

Electronic health records (EHRs) are ubiquitous in healthcare settings. EHR-derived clinical data can be used to improve identification and management of patients at-risk for or diagnosed with diseases; however, the quality of the data must first be assessed.1 Chemotherapy-induced nausea and vomiting (CINV) is a common cancer treatment-related symptom. The quality of the EHR documentation, specifically related to completeness, concordance, and heterogeneity,2 related to CINV assessment is not known. This study aims to evaluate the documentation practices of providers and the quality of the data regarding assessment and identification of CINV, and to develop a post-coordinated approach to better identify patients.

Methods

We evaluated the EHR of patients ≤ 18 years old with cancer who had received emetogenic chemotherapy in the outpatient setting between 1/1/2016 and 12/31/2017. The patients’ EHR data, including structured and unstructured data, was explored and annotated to identify domains where CINV was assessed and diagnosed. We extracted the variables of interest (Table 1) and compared the concordance of CINV-related documentation by prescriber (physician or nurse practitioner), RN, and ICD-10 billing codes. We then manually mapped the related terms and concepts for CINV, identified through our data extraction, through UMLS Metathesaurus to SNOMED-CT. We identified all terms, or atoms, listed in UMLS that were classified under nausea, vomiting, or chemotherapy-induced nausea and/or vomiting. Qualifiers and modifiers that provide a deeper level of granularity about temporality, severity, and negation were also searched and included. In an iterative process, we developed a concept mapping of CINV to inform a post-coordinated approach that would improve identification of patients with CINV.

Results

Thirty unique patients were included in the analysis. RNs were more likely to document an assessment of CINV compared with prescribers, and there was frequent discordance between RN and clinician reporting of CINV. ICD-10 codes designated by prescribers were concordant with primary oncologic diagnosis, but not chemotherapy encounters (47% agreement) or diagnosis of CINV (33% agreement with prescriber documentation). Mapping the identified terms through UMLs informed a post-coordinated expression that will increase identification of the relevant components to completely and accurately document CINV (Fig 1).

Conclusion

Inconsistencies by provider type and lack of completeness were noted in this pilot study assessing data quality of CINV documentation. Post-coordinated expression can improve the characterization of CINV specifically relating to temporality, severity, and negation.

References


Table 1: Patient and documentation characteristics

<table>
<thead>
<tr>
<th>Encounter characteristics</th>
<th>Total encounters N=30</th>
<th>ICD-10 Concordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary oncologic diagnosis</td>
<td>Leukemia (n=15)</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Lymphoma (n=6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Solid tumors (n=9)</td>
<td></td>
</tr>
<tr>
<td>Type of visit</td>
<td>Highly-emetogenic chemotherapy (n=8)</td>
<td>47%</td>
</tr>
<tr>
<td></td>
<td>Moderately-emetogenic chemotherapy (n=20)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Follow-up from prior chemotherapy (n=2)</td>
<td></td>
</tr>
<tr>
<td>Clinician assessed CINV</td>
<td>Yes (n=22)</td>
<td>N/A</td>
</tr>
<tr>
<td>Clinician diagnosed CINV</td>
<td>Yes (n=6)</td>
<td>33%</td>
</tr>
<tr>
<td>RN assessed CINV</td>
<td>Yes (n=30)</td>
<td></td>
</tr>
<tr>
<td>RN reports CINV</td>
<td>Yes (n=10)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Mapping CINV specifically relating to

![Figure 1: Mapping CINV through UMLS](image-url)
Diagnosis Combinations Associated with Longer Length of Stay

Patrick E. Beeler, MD¹, Karol Tarcak, MSc¹, Ulrike Held, PhD², Edouard Battegay, MD¹
¹Department of Internal Medicine, University Hospital Zurich & University of Zurich, Switzerland; ²Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland

Introduction: Various factors such as specific diagnoses have been found to be associated with worse clinical outcomes among inpatients. However, little is known on interactions between diagnoses, i.e. combinations of conditions with an excessive impact not explained by additive effects. The aim of this study was to identify interactions between diagnoses that are independently associated with increased length of stay (LOS).

Methods: In this exploratory study at a large tertiary care academic medical center, all 287,255 inpatient stays over an eight-year period were included. Professional coders added ICD-10 codes to all diagnoses as defined by the physicians in charge of the patients, independent of whether the patient received care for a specific diagnosis or whether it contributed to the billing. We grouped the diagnoses assigned to the stays into 1,595 distinct categories based on the first 3 digits of the ICD-10 codes. A Lasso regression was applied to identify the 483 most important columns using cross-validation over alphas as well as lamdas, afterwards α=1 was selected. We then ran 19,608 regression models, each combining two distinct diagnoses, with a support >10, to evaluate whether there was evidence for a potential interaction. In each of these models, the main effects of all considered diagnoses were included. The outcome variable was LOS (days), which was log-transformed due to its skewed distribution. Raw computation time, with cross-validation and all subsequent regression models, took two days on a SMP machine with parallelization at 18 threads. We controlled for multiple testing by applying the Bonferroni correction (p value cut-off at α/19,608=2.5E-6). A final regression model was run including all 483 diagnoses, age, and 2,519 statistically significant interactions. Analyses were conducted with the software R.

Results: Due to the Lasso regression, age was kept, however, sex was not. The final model included 483 diagnoses, age, and 2,519 interactions. Eating disorders were separately investigated (not presented) due to a known standard LOS of 3 months for these patients. Ultimately, a total of 774 interactions were statistically significant. Of those, the top five beta weights and most frequent (N) interactions were:

<table>
<thead>
<tr>
<th>Interactions with descriptions</th>
<th>N</th>
<th>Adj.beta</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>C71.R29 Malignant neoplasm of brain</td>
<td>Other [nervous and musculoskeletal] signs</td>
<td>58</td>
<td>1.44</td>
</tr>
<tr>
<td>R65.G97 Systemic Inflam. Response [...]</td>
<td>[...] disorders of nervous system [...]</td>
<td>16</td>
<td>1.41</td>
</tr>
<tr>
<td>A41.S42 Other sepsis</td>
<td>Fracture of shoulder and upper arm</td>
<td>32</td>
<td>1.39</td>
</tr>
<tr>
<td>S82.S36 Fracture of lower leg [...]</td>
<td>Injury of intra-abdominal organs</td>
<td>74</td>
<td>1.22</td>
</tr>
<tr>
<td>Z37.O44 Outcome of delivery</td>
<td>Placenta praevia</td>
<td>316</td>
<td>1.20</td>
</tr>
<tr>
<td>O34.O60 [...] abnormality of pelvic organs</td>
<td>Preterm labour and delivery</td>
<td>1697</td>
<td>0.88</td>
</tr>
<tr>
<td>D62.F05 Acute posthaemorrhagic anaemia</td>
<td>Delirium, not induced by [...] substances</td>
<td>1605</td>
<td>0.85</td>
</tr>
<tr>
<td>C92.D61 Myeloid leukaemia</td>
<td>Other aplastic anaemias</td>
<td>866</td>
<td>0.84</td>
</tr>
<tr>
<td>I67.I60 Other cerebrovascular diseases</td>
<td>Subarachnoid haemorrhage</td>
<td>472</td>
<td>1.05</td>
</tr>
<tr>
<td>A41.D61 Other sepsis</td>
<td>Other aplastic anaemias</td>
<td>356</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Discussion: The prevalence of multimorbidity is increasing with a growing elderly population. Although many single risk factors worsening clinical outcomes have been identified as being critical, only very little is known about interactions of diagnoses that prolong LOS beyond additive effects due to the burden of the diseases. The negative impact of diagnosis interactions on the LOS of multimorbid inpatients has not been established yet and might have been underestimated so far. Finally, another crucial clinical outcome, 30-day readmissions, will be investigated in a next step by applying the same approach with multivariable logistic regression models.
FURTHER DATA VISUALIZATION DESIGNS TO SUPPORT SHARED DECISION MAKING ABOUT BLOOD PRESSURE

Jeffery L. Belden MD, Pete Wegier PhD, Shannon M. Canfield MPH, Sonal J. Patil MD, Victoria A. Shaffer PhD, Linsey M. Steege PhD, Mihail Popescu PhD, KD Valentine MS, Akshay Jain MS, Michael L. LeFevre MD MSPH, Richelle J. Koopman MD MS

1 Department of Family & Community Medicine, University of Missouri, Columbia, MO
2 Lumenfeld-Tennenbaum Research Institute, Sinai Health System, Toronto, ON
3 Department of Psychological Science, University of Missouri, Columbia, MO
4 School of Nursing, University of Wisconsin-Madison, Madison, WI
5 Department of Health Management & Informatics, University of Missouri, Columbia, MO
6 Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO
7 The Health Decision Sciences Center, Massachusetts General Hospital, Boston, MA

Abstract

We present a simplified data visualization of home and clinic blood pressures (BP) to support shared decision-making by the patient-physician dyad, visually emphasizing elements that have long term clinical significance (cardiovascular outcomes). We manage large volumes of patient-generated BP data by summarizing and smoothing to simplify interpretation of control and to remove distracting variability of no clinical consequence.

Background

Only about half of the 78 million US adults diagnosed with hypertension have their blood pressure controlled. The majority of BP measurements are currently gathered in physicians’ offices. Recent research has identified the important role of home BP measurements, considered equal—or even superior—to clinic BP measurements in their predictive value. Home BP monitoring is becoming more common, with seamless uploading of home BP measurements directly into a patient’s EHR on the horizon. As with recent innovations in continuous personal heart rate monitoring (e.g., Apple Watch), a future of continuous BP monitoring is rapidly approaching. While past literature has shown that data visualization of physiological data is highly effective at improving clinical understanding, we must design for the transition from the current landscape of BP data scarcity to a future of data abundance and, if we are not prepared, data overload.

Objective

Current EMR systems have only limited abilities to visualize BP data. Our goal was to develop a display—simple enough for patients but powerful enough for physicians—to allow for the visualization of a patient’s home BP measurements to support shared decision making.

Methods

We used a rapid prototyping process in which candidate visualizations were iteratively refined based on regular feedback from both patients and physicians. Several focus groups with hypertensive patients, and family and internal medicine physicians, were held to gather feedback on candidate visualizations.

Results

Data density

With an abundance of home BP values, we can manage the volume of displayed data to enhance readability by smoothing, averaging, and/or hiding (binning multiple measurements into one displayed measurement; Figure, note 1). We find that we suppress display of single outliers when variability is high (when it’s likely unimportant clinically), but still display clusters of outliers (of likely or potential clinical consequence—e.g., a cluster of outliers might show symptomatic temporary low BP that requires clinical attention (Figure, note 2).

Smoothing

In experimental studies, we have previously found that patients perceive increased BP variability (Figure, note 3) as decreased control—despite the literature showing mean BP to be most predictive of health outcomes, with variability accounting for only a tiny fraction of the variance. To improve perception of BP control, we smooth raw data using the LOWESS algorithm (Figure, note 4).

Conclusions

The purpose of the visualization drives the design requirements. We designed for simplicity to support shared decision-making by the patient-physician dyad, visually emphasizing elements that have long term clinical significance (cardiovascular outcomes). Managing data density by binning in our display makes it more readable without hiding important outliers. Smoothing makes it easier to identify the recent average trend, reducing the recent distraction of any variability.

References

OncoMX: an Integrated Cancer Mutation and Expression Knowledgebase for Biomarker Evaluation and Discovery

Amanda Bell¹, Raja Mazumder¹, Daniel Crichton², K. Vijay-Shanker³, Frédéric Bastian⁴,⁵, Hayley Dingerdissen¹, Samir Gupta³, Evan Holmes¹, Robel Kahsay¹, Heather Kincaid², David Liu², ASM Ashique Mahmood³, Marc Robinson-Rechavi⁴,⁵, Stephanie Singleton¹

¹Department of Biochemistry and Molecular Medicine, The George Washington University, Washington, DC, USA; ²Jet Propulsion Laboratory, NASA, Pasadena, CA, USA; ³University of Delaware, Newark, DE, USA; ⁴Swiss Institute of Bioinformatics, Lausanne, Switzerland; ⁵University of Lausanne, Lausanne, Switzerland

Description
Cancer genomics studies produce massive, heterogeneous datasets with variable characteristics, creating challenges during data integration, analysis, interpretation of results, and extrapolation of meaningful observations across different datasets. Consequently, attempting to utilize, analyze, or combine datasets from multiple sources proves inefficient and convoluted. The OncoMX knowledgebase and corresponding web portal is actively being developed to mitigate such challenges by leveraging biomedical ontologies to unify disparate data while emphasizing the availability of multi-faceted evidence for cancer biomarkers. The OncoMX knowledgebase is accessible via a searchable web portal and enables integrative exploration of cancer genomics data from mutation, disease and normal expression, literature mining, and biomarker and pathway databases. OncoMX was designed, with key user input, to better facilitate cancer biomarker research and discovery through four major use cases: 1) exploration of cancer biomarkers, 2) evaluation of mutation and expression in an evolutionary context, 3) side-by-side exploration of published literature for mutation and expression in cancer, 4) and exploration of a specific gene or biomarker within a pathway context. OncoMX is expected to benefit basic cancer research, promote efficient consumption of information by end users, and ultimately improve biomarker and cancer detection capabilities. OncoMX is available at https://www.oncomx.org/.

Figure 1. DNA/RNA-seq-based mutation and expression data are integrated and unified by DO and Uberon terms into BioMuta and BioXpress at the HIVe Lab. Data is supplemented by Bgee normal expression, literature mining for mutation (DiMeX) and expression (DEXTER) in cancer, functional annotations, EDRN biomarkers and a pilot dataset of FDA Approved Breast Cancer biomarkers, and Reactome pathways. All data are subject to unification processes and provenance capture before being consumed by the OncoMX processing pipeline.
Providing Data Security Guidance for Researchers

Douglas S. Bell, MD PhD\textsuperscript{1,2}, Spencer Soohoo, PhD\textsuperscript{1,3}, Alex Bui, PhD\textsuperscript{1,2}, Ann S. Chang\textsuperscript{1,2}, Marianne Zachariah\textsuperscript{1,2}, Ross Fleischman MD MS\textsuperscript{1,4}, Liz Y. Chen, MBA PhD\textsuperscript{1,4}, Omolola Ogunyemi PhD\textsuperscript{1,5}, Robert A. Jenders, MD MS\textsuperscript{1,5}

\textsuperscript{1}UCLA Clinical & Translational Science Institute, Los Angeles, CA; \textsuperscript{2}UCLA David Geffen School of Medicine, Los Angeles, CA; \textsuperscript{3}Cedars-Sinai Health System, Los Angeles, CA; \textsuperscript{4}Los Angeles Biomedical Institute at Harbor UCLA Medical Center, Torrance, CA; \textsuperscript{5}Charles R. Drew University of Medicine and Science, Los Angeles, CA.

Introduction

When investigators undertake research, they assume the responsibility to protect subjects’ privacy in collecting, transmitting, analyzing and storing health-related information. However, researchers are often uncertain how these responsibilities play out across the different types of data that they deal with and across steps in the data lifecycle. This uncertainty may inhibit their research activities, or conversely, may put them and their institutions at risk. We set out to create a guidance document for researchers to serve as a reference for their data security responsibilities.

Methods

The UCLA CTSI convened a Data Security Work Group (DSWG), consisting of the Informatics Program leads and the information security officers from each CTSI partner site, among others. The DSWG first evaluated several information security frameworks and chose the NIST 800-171 standard\textsuperscript{1} as the best framework for organizing an analysis of security practices. In addition, the working group determined that the sites’ highest priority was to provide data security guidance for researchers to use in preparing their protocols and data security plans. In response, the CTSI staff worked with security experts from each campus to create a grid that guides researchers about their minimum responsibilities for each type of data they may collect and across each stage in the data life cycle.

Results

We produced a Research Data Security Grid that provides investigators with guidance for protecting 4 different types of data across 10 steps in the data life cycle. The four data types are: identified data derived from patient care (PHI), de-identified patient data, identified data collected from research subjects (which is termed research health information, or RHI), and Intellectual property (IP). The steps in the data life cycle are: data collection or generation, data transmission inside the institution, data storage by the researcher, data analysis, user training, audit and accountability, data transmission outside the institution, data storage and analysis at collaborating institutions, data destruction, archiving final data and sharing with outside researchers or the public. Note that simple use cases, such as a researcher collecting and analyzing data on her or his own, may involve only selected data types and may not involve all of the steps. Multi-institutional studies typically involve all of the steps. The grid also contains a glossary of terms and contains footnotes to indicate site-specific considerations. As an example, the guidance cell for data transmission of PHI reads “Encrypt data in transit. Use approved file sharing or file transfer platform.[2] Encrypted, password-protected email attachments are also permitted for internal data transfers.” The poster presentation will include the full grid and qualitative feedback on its acceptance among the UCLA CTSI institutions.

Conclusion

Data security guidance for researchers can by summarized in a grid format, organized along the axes of type of data being dealt with and stages in the data life cycle. We anticipate retaining this format as technology, and data security requirements, continue to evolve in the future.

References

Posda: An Open Source Curation Tool for Validation, Editing, and Inspection of DICOM Objects for Large Scale Imaging Archives

Bill Bennett, B.A.¹, Kirk E. Smith, B.S.¹, Quasar Jarosz, B.S.¹, Sonya Utecht, B.S.¹, Fred W. Prior, Ph.D.¹

¹Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR

Introduction

Disclosure of protected health information (PHI), data quality and provenance tracking are primary concerns of large scale publicly accessible image archives such as the National Cancer Institute’s Cancer Imaging archive (TCIA)¹². Posda³ is a set of open-source scalable curation tools that we developed to address these concerns.

Methods and Materials

Posda is a custom set of open-source curation workflow tools dockerized for easy deployment in a Linux environment. De-identification software such as RSNA’s Clinical Trials Processor (CTP) redact PHI assuming the DICOM conforms to a standard which is rarely the case, do not test for DICOM consistency or validity and are ill suited to find and redact PHI or correct errors on a large scale. Posda is a role-based application that parses DICOM tags and stores values into a database (even if they are not DICOM compliant), which allows DICOM encoding errors to be captured, analyzed, and corrected prior to publication. Posda maintains a DICOM validation rule set and maintains an up-to-date DICOM private tag dictionary. These underpin DICOM validation and guide de-identification processes. The original files are stored in a file system and subsequent modifications are stored in the Posda database.

<table>
<thead>
<tr>
<th>The current Posda release supports:</th>
<th>6. Visual inspection of RT object alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DICOM send and receive</td>
<td>7. Checks for duplicate SOP Instance UIDs</td>
</tr>
<tr>
<td>2. Summary spreadsheets for data verification</td>
<td>8. Semi-automated detection/correction of PHI</td>
</tr>
<tr>
<td>5. Visual inspection of DICOM images</td>
<td>11. Prioritization of multiple data streams</td>
</tr>
</tbody>
</table>

Results

Posda has been successfully used as the primary curation tool for TCIA for the past three years. A scalable TCIA workflow supports the curation of DICOM objects prior to release. Submissions Managers use Posda generated reports to ensure all DICOM objects are received from submitting sites, to schedule PHI reviews, check adherence to the DICOM standard, check linkages to referenced objects and review issues related to PHI. Data Support Specialists review PHI reports and images for potential PHI and report potential issues. DICOM curators edit the data to ensure DICOM objects are valid. TCIA’s Posda instance has processed over 100 submissions containing more than 57 thousand series and over 5 million images during that time. There have been 310 distinct edit operations affecting 23,406,766 files.

Conclusion

Posda is the primary curation tool used by TCIA to ensure submitted DICOM objects are free from PHI, are valid DICOM objects, and are scientifically useful. Open source Posda tools allow review and correction of DICOM objects for improved DICOM conformance and consistency within healthcare environments and image archives.

References

Potential Benefits of Dental Visits in Managing HbA1C And Identifying Diabetes and Prediabetes in the U.S.
Munder Ben-Omran, BDS, MS1, Paul Fontelo, MD, MPH2, Gabriela Lopez Mitnik, MPhil1, Darien Weatherspoon, DDS, MPH1, Bruce A. Dye, DDS, MPH1.
1 National Institute of Dental and Craniofacial Research, Bethesda, Maryland; 2 National Library of Medicine, Bethesda, Maryland

Introduction:
A Patient-Centered Healthcare delivery model can enhance disease identification and is appropriate for disease prevention and treatment delivery. Periodontitis is a chronic oral disease and given the bidirectional relationship between diabetes/prediabetes and periodontitis, timely referral of patients with diabetes and prediabetes during dental and medical encounters would benefit both patient and health care systems. The large number of people with diabetes or prediabetes in U.S. necessitates attentive monitoring and referral to decrease morbidity and associated financial burden of these conditions, and to improve health outcomes. The objective of this study was to determine how many adults with elevated (unmanaged) blood glycemic levels and periodontitis could potentially benefit from proper referrals in a fully integrated health-system.

Methodology:
Data from 6,691 adults aged 30 and older who participated in NHANES from 2011-2014 were analyzed to assess the relationship between managed and unmanaged HbA1C levels according to recent medical (MV) and dental visit (DV) utilization, after controlling for sociodemographic, health, and medical insurance factors. Unmanaged HbA1C population (HbA1C>= 5.7%) were determined using the American Diabetes Association guidelines. All statistical analyses were weighted and population-adjusted using SAS statistical software 9.4.

Results:
Information from study participants representing 143 million (M) non-institutionalized US adults aged 30 and over was analyzed. More than half of this population had unmanaged HbA1C levels (62.1 M). There were (2.4 M) undiagnosed diabetics, (39.3 M) undiagnosed prediabetics, and (53.5 M) adults with periodontitis. Among the population with a MV in the past year but no DV, there were (9.4 M) diagnosed diabetics, (6.1 M) diagnosed prediabetics, (1 M) undiagnosed diabetics and (22.7 M) undiagnosed prediabetics. The number of adults estimated to have periodontitis and a medical visit but no dental visit in the past year was (17.7 M). Among the population with a DV in the past year but no MV, there were (0.25 M) diagnosed diabetics, (0.34 M) diagnosed prediabetics, (0.36 M) undiagnosed diabetics and (2.1 M) undiagnosed prediabetics. Adults with periodontitis were more than twice as likely to have unmanaged HbA1C compared to individuals without periodontitis (OR=2.30, p>0.0001). Unmanaged HbA1C levels and periodontitis were more pronounced among older adults (age > 65 y), (affecting 17.2 M and 13.7 M respectively), and non-Hispanic blacks (5.7 M and 7.7 M respectively). Adults with a recent MV but no DV were 35% more likely to have unmanaged HbA1C compared to adults with recent MV and DV, controlling for sociodemographic factors and medical insurance.

Conclusion:
In a fully integrated health care system, upwards of (2.8 M) people with unmanaged HbA1C levels could be referred by dental providers to medical providers for further evaluation and (17.7 M) adults with periodontitis could be referred by medical providers to dental providers for further evaluation of periodontitis. Improving processes that facilitate referrals that lead to improvements in glycemic control and management of periodontitis could potentially reduce disease and economic burden in large health care delivery systems. Our findings suggest that tools to assist healthcare providers, such as machine learning algorithms and Clinical Decision Support Systems (CDSSs), operating with integrated electronic dental and medical records, could help primary health care providers (medical and dental) to: identify undiagnosed diabetic, prediabetic and periodontitis patients; facilitate referrals; and improve disease management.
Analysis of Representative Cases of Diagnostic Error in the Inpatient Setting
Kerrin Bersani^1, Kevin Carr BSIE^1, Nicholas Piniella^1, Kumiko Schnock, PhD, RN^1, Marc Pimentel, MD^1, Jacqueline Griffin, PhD^2, David W Bates MD, MS^1, Anuj K Dalal, MD^1

^1Brigham and Women’s Hospital, Boston, MA; ^2Northeastern University, Boston, MA

Introduction: Diagnostic errors is an ongoing threat to patient safety: an incorrect or delayed diagnosis can lead to patient harm, affecting up to 1 in 20 U.S. adults each year. Patients who encounter a diagnostic error may undergo additional testing, a delay in treatment, and potential harm due to inappropriate interventions or treatment. In the hospital setting, the root causes of failures in the diagnostic process have not been well characterized. In this study we describe our approach to identifying root causes of failures in the diagnostic process through analysis of representative cases of diagnostic error.

Methods: We conducted this study at Brigham and Women’s Hospital (BWH) in Boston, MA. We have identified the types of symptoms (e.g., chest pain) and conditions (e.g., pneumonia) in which diagnostic error may lead to harm for patients admitted to the general medicine service at BWH. We retrospectively identified cases of potential diagnostic error from various institutional forums and reporting structures (Table 1). To classify and prioritize root causes of error, we have modified the Safer Dx framework and the DEER Taxonomy tool to correlate failure points to diagnostic process dimensions (Figure 1), which we are using to confirm diagnostic error.

Results: To date, we have identified 1611 reports from various reporting structures (Table 1) that might be associated diagnostic error. We have currently confirmed errors in 8 cases using our chart review tool.

Conclusion: Identifying cases of diagnostic error is a labor-intensive process that requires retrieving potential cases and then confirming the presence of errors using a validated instrument. Our team is conducting fault tree analyses of confirmed cases of diagnostic error to prioritize failure points in the diagnostic process during acute care. We will then aggregate fault trees across all representative cases to identify common root causes of diagnostic error. Failure points will be prioritized to serve as potential targets of technological interventions to prevent diagnostic error.

Figure 1. Failure Points Mapped to Diagnostic Process Dimensions

<table>
<thead>
<tr>
<th>Process Dimension</th>
<th>Failure Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient-provider encounter and initial assessment</td>
<td>Failure or delay in ordering needed test(s)</td>
</tr>
<tr>
<td>Diagnostic test performance and interpretation</td>
<td>Failure or delay in performing needed test(s)</td>
</tr>
<tr>
<td>Follow-up and tracking of diagnostic information</td>
<td>Suboptimal test sequencing</td>
</tr>
<tr>
<td>Subspeciality consultation</td>
<td>Failure to order correct test(s)</td>
</tr>
<tr>
<td>Healthcare Team Communication and Collaboration</td>
<td>Failure to order test(s) in correct way</td>
</tr>
<tr>
<td></td>
<td>Identification failure (e.g., sample mix-up, mislabeled specimen)</td>
</tr>
<tr>
<td></td>
<td>Technical or processing error (equipment problem, skill issue)</td>
</tr>
<tr>
<td></td>
<td>Specimen delivery problem (e.g., specimen never sent, lost specimen)</td>
</tr>
<tr>
<td></td>
<td>Erroneous reading of test (lab/radiology)</td>
</tr>
<tr>
<td></td>
<td>Erroneous clinician interpretation of test</td>
</tr>
</tbody>
</table>

Table 1. Identification of Representative Cases

<table>
<thead>
<tr>
<th>Reporting Forum</th>
<th>Description</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital Medicine QA Committee</td>
<td>Presentation of cases to hospitalist physician colleagues where a known error has occurred, and discuss causes of error and future prevention</td>
<td>• 8 cases where a known diagnostic error has been presented</td>
</tr>
<tr>
<td>Morbidity &amp; Mortality Review</td>
<td>Information on patients who died during hospitalization, rated by preventability (1-‘Least Preventable’ to 5-‘Most Preventable’)</td>
<td>• Of 358 cases, 158 (34.9%) had a preventability 3-5, or free-text comment</td>
</tr>
<tr>
<td>RL Solutions Safety Reporting</td>
<td>Self-reported cases by hospital staff regarding patient safety concerns, ranked by severity (0: ‘No Harm, did not reach the patient’ to 4: ‘Death’)</td>
<td>• Of 4225 reports, 1400 (33.1%) had a severity score of 1 or higher</td>
</tr>
<tr>
<td>Patient and Family Relations Reports</td>
<td>Reports by patients and their families on dissatisfied Care/Treatment they received at the hospital</td>
<td>• Of 119 reports, 45 (47.8%) related to diagnosis or care</td>
</tr>
</tbody>
</table>

Acknowledgements: The PSLL project is supported by funding from AHRQ (R18-HS026613).

Reference:

The Role of Bipartite Networks in Stratified Predictive Modeling
Suresh K. Bhavnani PhD1, Clark Andersen MS2, Yu-Li Lin MS2, Emmanuel Santillana MS1, Tianlong Chen PhD1, Yong-Fang Kuo PhD3

1Instit. for Translational Sciences, 2Preventive Medicine and Community Health, Univ. of Texas Medical Branch

Abstract
Although several studies have identified strong heterogeneity in comorbidity profiles of patients readmitted to the hospital, current CMS models do not use such information. Here we used bipartite networks to automatically identify biclusters of patients and comorbidities, and used them to develop stratified predictive models, with implications for precision medicine.

Introduction
A wide range of studies on topics ranging from molecular to environmental determinants of health have shown that most humans tend to share key characteristics (e.g., comorbidities or genes) forming distinct patient subgroups. A primary goal of precision medicine is to identify such patient subgroups and to infer their underlying disease processes in order to design interventions that are targeted to those processes. To address this goal, here we use bipartite networks to identify clinically meaningful subgroups of readmitted patients with congestive heart failure (CHF) from the Medicare dataset, and explore the use of those subgroups to develop stratified regression models for predicting hospital readmission.

Method and Results

Data: We extracted cases (CHF patients readmitted within 30 days of hospital discharge) and controls (not readmitted within 90 days of discharge, matched by age, gender, race, and Medicaid eligibility as a proxy for economic status) from the Medicare Provider and Analysis Review (MedPAR) Files for 2013 and 2014 combined. This resulted in a training dataset (25,775 matched pairs), and a replication dataset (25,775 matched pairs). Variables consisted of 42 unique comorbidities highly prevalent in the elderly extracted from the Elixhauser, Charlson, and CMS-CC indices, in addition to history of coronary artery bypass graft surgery, critical in the analysis of CHF readmission. Using odds ratio with Bonferroni correction, we identified 37 comorbidities that were both significant in the training dataset, and replicated in the replication dataset.

Method: The analysis consisted of three modeling stages: 1. Visual Analytics Modeling was done by (a) representing the data as a bipartite network where nodes represented either patients or comorbidities, and the edges represented the presence or absence of a comorbidity, (b) identifying patient subgroups and their most frequently co-occurring comorbidities using bicluster modularity (Q) and testing its significance through comparisons to 1000 random permutations of the data, and (c) using the Rand Index (RI) to measure the similarity of comorbidity co-occurrence between the training and replication datasets, and testing the RI significance. The network was presented to a geriatrician specializing in CHF hospital readmission, who examined the clinical meaningfulness of the co-clusters in the network. 2. Classification Modeling was done by (a) building and validating a multinomial regression model to predict the probability of a patient belonging to each of the subgroups identified by the visual analytics based on their profile of 37 comorbidities described above, and (b) classifying each patient to a subgroup based on the highest probability of cluster membership. 3. Predictive Modeling was done by (a) building and validating a logistic regression model using all the patients lumped together to predict the probability of a patient being readmitted based on their profile of 37 comorbidities, age, sex, and race, and (b) using the C-Statistic (a validation measure of predictive accuracy) to compare the above lumped model to the validated stratified logistic regression models representing each of the patient subgroups (identified from the visual analytics), using the same above variables.

Results: The visual analytics identified 4 biclusters with significant modularity (Q=0.17, z=8.69, p<0.001), and significant co-occurrence replication (RI=0.94, z=17.66, p<0.001). The pulmonologist verified that the patient subgroups were clinically meaningful (e.g., a patient subgroup had a combination of psychological comorbidities with COPD, known to be high risk for CHF readmission). Fig. 1 shows how the C-Statistic of the lumped model (which ignores patient subgroups) compares to the 4 stratified regression models. As shown, Subgroup Model-1 (SM-1) has a significantly lower C-Statistic compared to the lumped model, suggesting that the current approach (used by the Centers for Medicare and Medicaid Services) of using a single model for predicting CHF readmission, significantly underperforms for a patient subgroup. Our current work explores whether a similar heterogeneity in predictive accuracies exists in other conditions tracked by CMS such as stroke, with the goal of developing targeted predictive models that have equivalent and high predictive accuracy across all patient subgroups. Such models could help hospitals predict and design interventions targeted to specific patient subgroups.

Acknowledgements. Funded in part by PCORI #ME-1511-33194.

References

Fig. 1. The lumped model significantly underperforms for a subgroup of CHF patients.
Developing Synthetic VA Healthcare Data in OMOP CDM Model

Jiantao Bian, PhD¹, Hamid Saoudian, MS¹, Brett R. South, PhD¹, Kristine E. Lynch, PhD¹, Benjamin Viernes, MPH¹, Michael E. Matheny MD, MS, MPH², Scott L. DuVall, PhD¹
¹VA Salt Lake City Health Care System and University of Utah, Salt Lake City, UT, USA
²VA Tennessee Valley Healthcare System and Vanderbilt University, TN

Introduction

The field of Biomedical Informatics could greatly benefit from new sources of open, sharable data. However, legislative mandates under the Health Insurance Portability and Accountability Act (HIPAA) and the Common Rule that protect patient privacy, confidentiality, and consent strictly limit the availability of shareable patient healthcare data. An open area of research and application of informatics methods has focused on generating realistic synthetic healthcare datasets that are representative of target patient populations. Under most situations patient identifiable clinical data are “locked down” and subject to institutional and federal regulations regarding data sharing that include de-identification approaches that focus on removing the 18 types of PHI specified under the HIPAA safe harbor standard and anonymization which goes further to remove all linking information to minimize risk of patient re-identification. However, these methods are labor intensive and may still put the patient at risk for re-identification when data are publicly released. Our goal is to provide a realistic synthetic data resource, SynVA (Synthetic VA), that could be used for research, software development, and training by healthcare community.

Methods

We limited our work to only structured data obtained from the VA Corporate Data Warehouse (CDW) that had been previously transformed into OMOP Common Data Model Version 5.1. We focused on simulating patient records from outpatient clinical visits by modeling the relational dependencies and conditional probabilities of actual VA medical record data. The methodology consisted of two parts. First, we calculated the aggregated distribution of clinical data from real patients. We then populated the synthetic data according to these distributions. This distribution-based method enables us to generate any arbitrary size of synthetic patient population. We iteratively populated synthetic data into the OMOP Person and Condition Occurrence tables. For the Person table, we categorized our entire VA Patient population (11.3 million) into 200 demographic groups according to the joint distribution of patient gender (M/F), age (ten groups), race (five groups) and ethnicity (Hispanic or Latino/Not Hispanic or Latino). We selected 60 groups based on the inclusion criterion of at least 10,000 patients/group. This allowed us to include the majority of patients and also ensure each group had enough patients to generate its distributions for the clinical elements (i.e., clinical condition, medication, and lab). For the Condition Occurrence table, we selected the top 1,000 clinical conditions in each of the 60 patient groups to ensure adequate coverage of clinical conditions. We evaluated the resultant synthetic data by comparing the distributions of condition records between the synthetic patient data and the original patient data (1). All steps were implemented in SQL Server Management Studio.

Results

We successfully generated 1,000,000 synthetic patients in the Person table to match the distribution of source population with respect to demographics. The most common demographic group was patients male, 65-74 years old, white, and not Hispanic or Latino. We generated 145,507,660 records in the Condition Occurrence table. The proportion of condition records from each of the 60 groups ranged from 0.004% to 24.16%, which closely mirrored that of our source data. The differences in distribution between synthetic and real data ranged from 0.0003% to 0.28%.

Conclusion

A distribution-based method is promising for developing synthetic health care data that protects patient privacy and confidentiality. Future versions of SynVA will populate additional OMOP tables (e.g., Visit Occurrence, Drug Exposure), consider the temporal and clinical plausibility between data elements, and incorporate more source-to-SynVA validations (e.g., Charlson Comorbidity Index comparison).

References

Tracking Breast Cancer Survivorship over time using the Claim Framework

Catherine L. Blake, PhD
University of Illinois, Champaign, IL, USA

Introduction

Concerted efforts such as STEEP\(^1\), DATECAN\(^2\), and REMARK\(^3\) are underway to harmonize breast cancer outcomes. At the same time automated methods to detect outcomes from peer-reviewed literature have advanced from identifying sentences to identify a specific noun phrase\(^4,5\). We show how integrating outcome entities into the Claim Framework\(^6\) enables physicians to visualize how survivorship has changed over time for different treatment strategies.

Method

Searches were conducted for brand and generic names of five drugs that are employed in systemic treatments of breast cancer from the 2016 baseline MEDLINE files. The treatments were chemotherapy (Doxorubicin and Docetaxel), hormone therapies (Tamoxifen, Raloxifene, and Bazedoxifene), and a targeted therapy (Trastuzumab). Explicit claims that involved a survivorship outcome were extracted and quantified from the 88,727 abstracts. Explicit claims were further grouped into those that show an improvement (e.g. improve, extend), a change where directionality was not provided (e.g. change, effect), and where survivorship did not improve (e.g. reduce, decline).

Results and Conclusion

Survivorship outcomes were reported 65,614 times in 21,906 abstracts. There were 12,445 explicit claims reported in 8,850 abstracts. Figure 1 provides a visual summary of how survivorship outcomes have change over time. Even though these treatment strategies were introduced at different times (see the x-axis) the proportion of abstracts with improved survivorship generally peaks just after a strategy is first introduced. These results suggest that automated methods could be used to track changes in outcomes over time and potentially to identify promising new strategies as they emerge. Further work is needed to differentiate between clinically relevant and surrogate endpoints.

![Figure 1. Percentage of abstracts that report an increase (purple), change (grey), or decrease (orange) in breast cancer survivorship for chemotherapy, hormone therapy, and targeted therapies over time.](image)

References

Sequential Analysis of Handoffs in a Mixed Methods Study

Katherine Blondon, MD, PhD¹, Pejmon Noghrehchi², Brian Hilligoss, PhD², Mustafa Ozkaynak, PhD³

¹University Hospitals of Geneva, Geneva, Switzerland; ²University of Arizona, Tucson, AZ, USA; ³University of Colorado, Denver, CO, USA

Introduction

Handoffs are the transfer of accountability and patient information from one healthcare provider to another. Handoff-related errors are common and can threaten patient safety and quality of care. One approach to improve handoffs includes the use of mnemonics such as the SBAR (Situation, Background, Assessment and Recommendation)¹ to help organize the content of handoffs. Sequences matter for effective cognitive processing when communicating large amounts of information. We hypothesized that some handoff content may already be structured with components of the SBAR, thus we aimed to study to what extent the sequential organization of handoff communication content is congruent with SBAR to inform future training interventions.

Sequence analysis is a challenging analytic task. Current software tools to support qualitative analyses do not sufficiently support sequence analyses. They can facilitate coding, linking, mapping of the data and querying in large datasets, but lack visual or statistical tools for sequence analyses. We propose an approach combining two software applications to conduct the sequence analysis, and discuss the implications of our results for future handoff training.

Methods

This is a secondary analysis of simulation-based verbal sign-outs. Twenty-one internal and family medicine physicians from a teaching hospital were enrolled to study handoffs in a simulated nightshift setting. Each participant received the handoff of four fictitious patients, and were given access to the EHR. Participants then received a page about each patient from ward nurses, which they needed to manage. They signed out their four patients to an investigator. This study focuses on one of the eight patient cases from this sign-out dataset, which was transcribed verbatim, and coded in Atlas.ti using an in-depth thematic analysis. Codes were grouped into the SBAR themes. Each code group then was assigned a time stamp to preserve the temporal order of the codes that were identified in thematic analysis. The list of time-stamped code groups were analyzed using Eventflow² which provides aggregated summaries of sequential data patterns in an interactive visual way.

Results

EventFlow visualizations highlighted the sequence patterns of SBAR elements stated by the eleven participants for this patient case. Durations of sign-outs varied between 2:06 and 6:46 minutes. All participants communicated patient identification features in the beginning of the sign-out. Overall, the majority of the participants communicated Background information early in the sign-out, and the Assessment was followed by Recommendations. The Situation information, however, was spread out during the sign-out in multiple segments; one participant even ended with a Situation segment. Recommendations were also often in multiple segments, following either a Situation or Assessment segment. There was no association between sign-out duration and organization of information.

Conclusion

Our mixed methods approach provides insight into the information sequences used in sign-outs. Based on this analysis, training could focus on the Situation segment, which should be given earlier in the sign-out. On-going research is exploring the generalizability of these results with the other patient cases. Future research should focus on how technology and training interventions can help clinicians organize their sign-outs better, supporting the use of validated mnemonics that can reduce medical errors.

References

Providing a User-oriented Interface to Enable Semantic Queries in PRISM
Jonathan P. Bona PhD, Joseph Utecht BS, Mathias Brochhausen PhD
University of Arkansas for Medical Sciences, Little Rock, AR, USA

Introduction
The Platform for Imaging in Precision Medicine (PRISM) expands core capabilities of The Cancer Imaging Archive (TCIA) to support the evolving requirements of cancer Precision Medicine research. This poster highlights our work extending PRISM’s ability to integrate and manage clinical, patient demographic, and other non-image data that typically accompany many of the collections comprising millions of de-identified DICOM-format cancer research images currently stored and managed by TCIA. Non-image data provide valuable information about the subjects of these images useful for search, exploration, and use of these data. However, clinical and other non-image data have historically been submitted and stored as spreadsheets lacking a common representation scheme, and that are not easily combined or queried. We integrate these into a semantic knowledge graph in a triple store database, and are developing user-facing tools for exploration of these data, allowing users to search across image collections by patient demographics, tumor location and disease types, diagnoses in clinical data, and more.

Methods
This poster presents our representation, transformation, and use, of de-identified non-image data accompanying two different public TCIA collections of head and neck cancer images. Though these collections originate from different sources, some types of non-image data are found in both, including basic demographic information, tumor sites, cancer types, and HPV diagnoses. These elements have been the focus of our work to integrate data from these collections, translating them to a shared semantic representation that is used for similar data appearing in other collections. Our representations for these data use OBO Foundry ontologies, including the Human Disease Ontology (DO), and the Uber Anatomy Ontology (Uberon). Instances are generated for each occurrence of the data elements we use and linked to ontology classes that explicitly represent the entities that the data are about (anatomical locations, disease types, diagnoses, demographic characteristics, etc). We have written Python programs that apply the matching semantic representation patterns to the input data, generating OWL/RDF files as output. These are loaded into a triple store, resulting in a queryable knowledge graph with built-in logical inference. This is immediately useful, for instance to find non-exact but matching results based on partonomic knowledge in an anatomy ontology. A simple web interface leverages semantic web technology without requiring the user to understand the technical details, hiding the complexity of OWL/RDF representation, the details of the ontology classes, and the SPARQL queries used. This interface presents simple, familiar user interface elements, such as an auto-completing typeable select box for tumor sites that is populated using the human-friendly labels attached to Uberon classes used in the instance data; or input boxes allowing the user to specify a search range for the age of subjects in the collections. These search elements generate a combination of SPARQL queries in the triple store and SQL queries in the relational database that manages TCIA image collection metadata.

Results
The knowledge graph resulting from this transformation links patient identifiers to representations of patients and patient characteristics and disease and diagnosis information, and the relations among those. Ontology terms classify these instances, defining the classes and relations used. This graph can be queried using the interface we have developed to identify records matching criteria based on fields that were previously not queryable, for instance to find records across head-and-neck cancer collections for Male patients over the age of 55 with a positive HPV diagnosis and a primary tumor in the oropharynx. The search results produced by this interface link directly to downloadable images, allowing users to quickly identify and view images that match their selection criteria.

Discussion and Future Work
This work demonstrates the usefulness of ontologies and semantic web tools to transform non-image data accompanying cancer image collections into shared semantic representations that enable integration, query, discovery, and exploration of both image and non-image data in these collections. Work is ongoing to roll out semantic representations of additional data and data types, and to develop further user-facing tools to support exploration and use of these data.

Acknowledgement
This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Under this contract the University of Arkansas is funded by Leidos Biomedical Research subcontract 16X011. Funding was also provided by U24CA215109.
SMOREs: Synergistic Medication Organizer for RxNorm and Expansions, a Tool For Improving Semantic Relationships of Institutional Medications to Standardized Ontologies
Robert L Bradford, BS, CCIS, Marshall Clark, BS, Emily Pfaff, MS
University of North Carolina at Chapel Hill, Chapel Hill, NC

Introduction
Analyzing medication data longitudinally presents a significant challenge due to new medications routinely coming to market and others being removed. These changes in available medications lead to constant fluctuations in the available discrete codes used to describe those medications. While there are different coding sets, such as RxNorm, NDC, CPT or SNOMED-CT that provide semantic concepts for these medications, they must be kept in sync with the prescribe-able medications to maintain accuracy. Multiple endeavors at the University of North Carolina at Chapel Hill (UNC) require the semantic interoperability of these different medications codes across data sets. One such endeavor, the clinical research networks (CRN) including PCORnet and Accrual to Clinical Trials (ACT) have chosen to harmonize medication coding around RxNorm, a code set curated by the National Library of Medicine (NLM). In the development of a prototype clinical triplestore utilizing the FHIR® R3 RDF specification, we identified the need to develop a comprehensive and harmonized approach to mapping medication codes. Additionally, a key component of our clinical triplestore includes the integration of insurance claims and EMR data. Therefore, we developed the Synergistic Medication Organizer for RxNorm and Expansions (SMOREs) to input local medication codes and produce mappings to standardized ontologies such as RxNorm. SMOREs also aims to address gaps present in both EMR and claims medication data. For example, the Carolina Data Warehouse for Health (CDWH), contains EMR data from 2004, including medication prescriptions for which at least 21% were discovered to have no internal mappings to RxNorm, in contrast to the absence of mappings within the available insurance claims. By using SMOREs we were able to acquire semantic interoperability of EMR and claims medication data.

Development
The NLM and the Food and Drug Administration (FDA) have provided multiple API endpoints that provide access to different ontologies. To develop SMOREs, multiple endpoints were identified that could be used to lookup information on medication ontologies (Table 1). SMOREs was structured to allow for easy addition of new API endpoints as additional data needs were identified. Development started with utilizing the RxNAV API to focus on RxNorm base information, followed by support for openFDA and UMLS. These APIs populate the outputs of the SMOREs pipeline: a valid FHIR® JSON file or CSV with “synonyms” for all requested medications. SMOREs was written in Python 3.6 and designed to be system agnostic.

Conclusions
SMOREs provides a multitude of functions that allow for significant expansion of information on local medications (Figure 1). Using SMOREs we were able to create a comprehensive medication dictionary that provided RxNorm code mappings at the Semantic Clinical/Branded Drug (most specific) and the ingredient/multi-ingredient (most generic) levels in FHIR® R3 RDF format for prescriptions from the EMR and medication claims data. Accomplishing this allowed us to group medications together in a semantically consistent manner utilizing relationships of RxNorm to other ontologies (ATC, NDFRT, etc.) and integrate them into our clinical triplestore.

References
1. https://www.hl7.org/fhir/
3. https://open.fda.gov/
Nursing Informatics Pioneers Identify Influencers

Juliana J. Brixey, Susan K. Newbold, Dianna Dodd

School of Biomedical Informatics and Cizik School of Nursing, University of Texas Health Science Center at Houston, Houston, Texas, USA; Nursing Informatics Boot Camp, Franklin, Tennessee, USA; Cerner Corporation, Kansas City, Missouri, USA

Introduction
The American Medical Informatics Association (AMIA) Nursing Informatics History Project was established to record and maintain the history of nursing informatics primarily in the United States to 2000 [1]. At the project start, 143 individuals were designated as pioneers of which 33 were interviewed and videotaped by convenience. Pioneers were defined as innovators, trailblazers, or groundbreakers in some aspect of nursing informatics, first to open a new area and prepare a way for other nurses to follow [2, 3]. The designated pioneers advanced nursing informatics through innovation in clinical practice, education, administration, and research. This study is an extension of a systematic review used to identify manuscripts published by the pioneers in PubMed for the years 2010-2015 as evidence of continued contributions to nursing informatics [4]. Contributions to the field of nursing informatics as evidenced by participation in social media – specifically Twitter and LinkedIn was studied. These applications were deemed mostly likely to be utilized for a public professional presence [5].

Methods
A retrospective review of publicly available transcripts available from the AMIA website was conducted for 33 American Pioneers in Nursing Informatics. Each transcript was read line by line to identify themes and specific individuals as identified by pioneers as impacting nursing informatics. A concept map was created for each pioneer to visualize the themes and identified individuals.

Results
The following themes were identified
- Impacts nursing informatics
- Contributes to nursing informatics
- Brings the practice along
- Guiding lights
- Pioneers
- Leading edge
- Pioneers as friends

Examples of concept maps illustrating the connections among pioneers.

Conclusions
This project is a continued systematic study of pioneers in nursing informatics. The text transcripts are a rich historic repository of information provided by the pioneers regarding their lived experience in nursing informatics.

Results from this study led to the identification of seven themes. These themes reveal how the pioneers described relationships as well as those influencing the nursing informatics. Relationships among the pioneers extended to nursing informaticians in other countries. This provides insight that nursing informatics was a global phenomena.

Further study is suggested to refine the themes identified in this study. It is recommended that future research explores connections formed from attendance at professional conferences and meetings.

References

Address for Correspondence:
Susan K Newbold. sknewbold@comcast.net
Estimating Uncertainty in Deep Image Classification

Katherine E. Brown, BS, Douglas A. Talbert, PhD
Tennessee Tech University, Cookeville, TN

Introduction

Recently, the application of deep learning to medical diagnosis has been an active area of research. Convolutional neural networks (CNNs) have made the computational processing of medical images more achievable. CNNs automatically extract important features from images that are then processed by an Artificial Neural Network to produce a prediction. Deploying CNNs in a clinical setting motivates a measure of the certainty of its prediction. In this study, we evaluate the effectiveness of various uncertainty measures on binary classification by CNNs.

Background

Gal et al.\textsuperscript{1} show uncertainty can be measured by calculating the standard deviation of the probability distribution resulting from repeatedly sampling a CNN with dropout active during training and testing. Leibig et al.\textsuperscript{2} demonstrate that this approach performed well in the classification of retinal images on the presence or absence of diabetic retinopathy. Tumor-Infiltrating Lymphocytes (TILs)\textsuperscript{3} are the infiltration of the immune system into a cancerous tumor and can be computationally detected in pathology slide patches using CNNs. In most patients, the presence of TILs is correlated with more positive response to treatment and increased survival.

Methods

We train two CNNs on a dataset of 85,000 images to determine the presence or absence of Tumor-Infiltrating Lymphocytes. The first network (VGG-16 predictor) is fine-tuning a VGG-16 CNN with ImageNet weights. VGG-16 is an effective CNN for image classification and performed well with this dataset\textsuperscript{2}. Our second model (Bayesian predictor) is a VGG-16 model augmented with dropout layers placed in the fully-connected layers, to create a network suitable for utilizing the Bayesian uncertainty measure\textsuperscript{1}. Our experiments compare two measures of uncertainty in deep image classification. The first is the Bayesian technique outlined by Gal et al\textsuperscript{1}. The second, naïve measure (softmax) is the absolute value of the difference of the predicted probabilities of binary classes from a standard CNN.

Experimental Analysis and Results

To evaluate these uncertainty measures, we first remove the most uncertain images from the classifier and measure the resulting performance of the classifier. Second, we simulate the use of an experienced pathologist by replacing the most uncertain labels with the known correct labels. For the first experiment, the accuracy of the Bayesian predictor improved more when using the Bayesian uncertainty quantification (UQ) (+17.0%) than when using the softmax UQ (+1.5%). The accuracy of VGG-16 predictor had a larger net increase in accuracy using the softmax UQ (+8.0%) compared to Bayesian UQ (+3.5%). For the second experiment, the accuracy of the Bayesian predictor had larger net increase from the Bayesian UQ (+16.5%) compared to the softmax UQ (+10.0%). The VGG-16 predictor had a larger net increase from the softmax UQ (+11.1%) compared to the Bayesian UQ (+9.0%) 

Conclusion

From our analysis, the best increase in accuracy occurs when uncertainty is measured directly from the predicting model. This is in contrast to using the more theoretically grounded Bayesian technique for measuring uncertainty and performing predictions with another model. In the future, we intend to further analyze this phenomenon to better understand UQ in deep learning.

References

The Role of Perceived Legitimacy in Clinical Decision Support

Julian Brunner, PhD, MPH, Elizabeth M. Yano, PhD, MSPH, Emmeline Chuang, PhD, Jeremy Shelton, MD, MSHS

VA HSR&D Center for Healthcare Innovation Implementation & Improvement, Los Angeles, CA; University of California Los Angeles, Los Angeles, CA

Introduction

There is growing recognition that clinical decision support (CDS) does more than inform or remind: it affects provider behavior through complex psychosocial channels, most notably by signaling the legitimacy of a given clinical practice to providers and patients. However, the specific role that legitimacy plays in CDS remains unclear. The goal of this analysis is to develop a conceptual model of the role of perceived legitimacy in CDS. Our objective is to describe how perceived legitimacy works as a mechanism for CDS to shape provider behavior, and identify moderating factors that determine the role of perceived legitimacy.

Methods

We conducted semi-structured qualitative interviews about CDS with 25 primary care providers from VA medical centers and community-based outpatient clinics within the VA Greater Los Angeles healthcare system. We derived our model after conducting thematic analysis of interviews with deductive and inductive techniques.

Results

Our analyses, summarized in Figure 1, identified moderating factors that shape the role of perceived legitimacy: provider uncertainty is typically a prerequisite for CDS to change behavior, provider attitudes about their organization determine the persuasiveness of organizational endorsement, and the extent to which CDS confers legitimacy depends in part on its content and framing, e.g. citing reputable sources, or showing the patient-specific information that informs the recommendation. We also identified multiple forms of provider-perceived legitimacy: hierarchical authority (e.g. my bosses want me to follow this recommendation), expert endorsement (e.g. locally-respected experts endorse this recommendation), institutional support (e.g. my organization or government endorses this recommendation), and peer support (e.g. members of my in-group endorse this recommendation). Finally, we found that CDS can also shape patient-perceived legitimacy, e.g. when providers point to computerized recommendations to demonstrate their institutional/societal endorsement.

Discussion

Failure to understand the psychosocial mechanisms of CDS can lead to ineffective or counterproductive interventions. Clearly specifying the role of legitimacy in CDS may enable CDS design to better target this mechanism whose role is widely acknowledged but imprecisely understood.

References


Electronic Informed Consent (eIC) Platform for Clinical Trials: An Operational Model and Suite of Tools for Consent Authoring, Obtaining Informed Consent, and Managing Consent Documents

Michael Buckley, MS, MBA, Joseph Lengfellner, Matthew Koch, Hector Pacheco, Carol Hoidra, MDiv, Dorothy Damron, MS, Collette Houston, Roy Cambria, Ann Rodavitch, MA, Paul Sabbatini, MD
Memorial Sloan Kettering Cancer Center, New York, NY

Introduction
The informed consent process is the foundation of human research subject protection, and studies have shown that enhancing the consent experience with introductory videos and visual aids, can improve participant engagement and comprehension.1,2 The eIC platform was developed to augment educational alternatives for research participants, reduce administrative time and effort associated with paper consenting, improve the IC audit trail, and streamline consent document authoring for clinical research studies. This web-based platform is device-agnostic and browser-independent; it is now used by 35 Services for 36 institutional and sponsored therapeutic and non-therapeutic clinical trials. Three protocols in the platform have an educational video embedded in the eIC, and 5 have an embedded image flow that gives an overview of the protocol timeline for tests and clinic visits.

Methods
To evaluate the pros and cons of the eIC platform versus the paper, we assessed: 1) the availability of the finalized document in the electronic health record (EHR), 2) processing time, and 3) the accurate completion of required data fields. A five-question survey was used to assess participant’s feedback on the eIC process, and free text responses were reviewed for common topics. Participants drive the decision to use electronic or paper consenting, and the discussion process is the same for both methods.

Results
As of July 2019, 201 active consenting professionals use the module, and 5,320 research participants were consented. Average eIC monthly accrual from January-June 2019 was 485 (+/- 45). Compared with paper consent forms, which take ~ 72h to post to the EHR, the signed eIC is sent and stored in both the EHR and the Patient Portal in < 2 minutes. The eIC platform decreases administrative effort associated with paper-based consenting by 5-15 minutes/form. The eIC module has a robust audit trail that tracks the consent session and participant interactions via timestamps. We compared results of 170 patients consenting to one protocol during the same timeframe; 85 used eIC, and 85 used paper. eIC use increased the completion of required data fields in the consent form by 4%, versus paper. Surveys were sent to 2,120 eConsent users, with 456 responses received (22%). Most respondents (370, 81%) noted electronic consenting was very easy (168), or easy (202) to use. Only 14 respondents (3%) noted that electronic consenting was somewhat difficult or difficult to use, and 72 were neutral. The majority of respondents (444, 97%) noted they would recommend electronic consenting to another patient at MSK. Free text responses were submitted by 229 respondents (50%), and surfaced the consistent themes noting the electronic process was simple, convenient, and user friendly.

Conclusion
The eIC platform provides an innovative operational model for electronic consenting in the academic medical setting, enhances the engagement and education of study participants, and improves quality and compliance in the consent process. We plan to expand our preliminary analysis by creating an assessment tool that will gauge participant understanding/engagement in the consent process.

References
Combining Neural Nets

Nithin Buduma¹, Aviv Segev, PhD², Dorothy W. Curtis, M.Sc.¹
¹Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA; ²Department of Computer Science, University of South Alabama, AL, USA

Abstract

To date many papers have been published describing neural networks’ successes at classifying medical images. Our work seeks to combine these neural nets into a web service so that a client can submit a file (or files) and receive one or more analyses of their images. Our poster will show results from experiments testing the effectiveness of our combined neural nets on a variety of images. We see this as a step towards developing decision support software for medical image analysis.

Introduction

Patients as well as clinicians would usually like to consult multiple specialists when there is a tricky issue to resolve. Many papers have been published showing useful results from applying neural networks to medical images¹². We propose to bring these results to life by creating a web of neural nets which embody this knowledge. These neural nets will provide analyses to our clients of images they upload to our web service. To explore this architecture, we are first implementing our system to analyze images outside the medical domain. We are using the MNIST³ and CIFAR⁴ datasets. Our work will then progress to analyzing medical images.

System Description

Our Neural Net Web system consists of three main parts: (1) the user interface; (2) middleware; (3) and the individually trained neural nets. The user interface collects a file to be analyzed from the user. The middleware, or master neural net, uses a pre-trained convolutional net framework from torchvision⁵ to select individual neural nets should analyze the image. The middleware also converts the image if necessary to a common image format that all the individual neural nets have been modified to accept. The individual neural nets are built on top of a generic neural network which has been trained on a dataset collected from the web. When the selected neural nets have analyzed the image, their results are passed back through the middleware to the user.

Conclusion

We developed a model for decision support software for medical image analysis that integrates neural nets into an online system for image classification. Our software achieves 78% accuracy on the CIFAR-10 dataset and 99% accuracy on the MNIST dataset.

References

Feasibility of Using Computerized Adaptive Testing To Capture Patient Reported Outcomes in an Outpatient Setting: A Pilot Evaluation of PROMIS-CAT in Neurosurgery

Mohamad Bydon, MD; Kelsey Wolff, M.Ed.; Anshit Goyal, MBBS; Mohammed Ali Alvi, MBBS; Sandy Goncalves, M.S.; Travis Paul; Aaron Biedermann, MBA; Carolyn M. Macken, M.S.; Janine Kamath, MBA; Andrea Cheville, MD; Mark Nyman, MD; Mayo Clinic, Rochester, MN, United States

Introduction
Current challenges around Patient Reported Outcomes (PRO) administration include redundant and uncoordinated PRO collection leading to excessive patient burden and diminished response rates. In this pilot, we aimed to improve the completion rate of baseline Patient-Reported Outcomes Measurement Information System (PROMIS) questionnaires among patients presenting for outpatient evaluation in Neurologic Surgery, while reducing patient burden, using a computerized adaptive testing (CAT) format.

Methods
PROMIS-CAT was selected as the tool for capturing PROs using our institutional patient online services (POS) portal, linked directly to the electronic health record (EHR). Prior to the pilot, PROMIS-29 was administered online via an iPad at the point-of-care immediately prior to provider encounters, but was not immediately available within the EHR. The switch to PROMIS-CAT was driven by the benefits of enhanced efficiency and precision with CAT, as well as the immediate availability of PRO data in the EHR for provider reference. With PROMIS-CAT, participant responses inform a computer algorithm to select the subsequent items from an item bank that are most likely to inform trait estimation.

Results
Prior to the pilot, PROMIS-29 completion rate was 30%. During the two month pilot, 1863 patients were assigned PROMIS-CAT, of which 1285 or 69% of patients completed the questionnaire. The average number of questions answered by each patient was 45.2. Upper Extremity Function (which is not included as a domain within PROMIS-29) represented the domain with the highest number of questions administered while the least number of questions were administered in Fatigue and Physical Function domains.

Conclusion
EHR linked CAT may represent a valid tool to increase PRO collection rates. Selective domain administration may alleviate patient burden. Use of a cross-cutting PRO like PROMIS-CAT over multiple redundant legacy PRO measures, may also alleviate patient burden.
Developing a Performance Matrix for Multidisciplinary Teams

J. Callister, T. Kwan, S.K. Poon, The University of Sydney, Sydney NSW, AU
Y. Liu, L. Evans, P. Harnett, Western Sydney Local Health District, Sydney NSW, AU

Introduction
Multidisciplinary teams (MDTs) are considered the “gold standard” of care for patients with cancer. However, the performance of MDTs varies widely. While validated observational tools for assessing MDT performance have been developed, it is essential to develop a self-assessment tool for MDTs to review their performance and monitor improvement over time.

Aim
The aim of this study is to develop a reliable and valid performance matrix as an instrument for MDTs to evaluate their level of performance and monitor their improvement over time.

Methods
The methodology is shown diagrammatically below. Green = completed, Orange = current, Blue = future.

Phase 1: Criteria from an MDT member survey and key themes in the literature were incorporated into a 2-dimensional matrix. The matrix comprised five levels of performance (columns), from most basic to most advanced, and five components (rows) - governance and leadership, clinical discussion and decision making, meeting organisation, data analysis and research, infrastructure and workforce. Each component contained several sub-components.
Phase 2: The Delphi methodology was adopted to assess the reliability of subcomponents. 27 selected MDT members and Cancer Innovation Managers, with diverse medical, allied health and nursing expertise, were asked to categorise each criterion into a sub-component using a card sorting software. A formula was developed based on the concepts outlined in Bhattachery to evaluate the reliability and variance of members’ responses.
Phase 3: The card sorting test was repeated for the Performance levels and this was followed up by an online survey of a further 11 members using a 1 to 5 Likert scale.

Results
For the subcomponents, results showed 86% agreement between the opinions of MDT members and the original categorisation of criteria. 67% of these results were of high normal distribution. The results for performance levels showed general correlation but were not as consistent as those for the sub-components.

Discussion and Conclusions
This study aims to develop a valid and reliable performance matrix for MDTs to measure their performance against standardised criteria and monitor their improvement over time. Whilst preliminary results demonstrated good correlation for sub-components, levels of performance are currently being further reviewed by a small group as part of the Delphi process. Future research will focus on finalising performance levels and statistical validation of the matrix.

References

Acknowledgements
The following institutions are thanked for their support of this study: Western Sydney Local Health District, Sydney West-Translational Cancer Research Centre, and University of Sydney.
Investigating Pregnancy-Related Health Outcomes Among Patients with Sickle Cell Disease and Linking with Health Disparities
Silvia P. Canelón, PhD1, Mary Regina Boland, MA, MPhil, PhD1,2
1University of Pennsylvania, Philadelphia, PA, USA
2Children’s Hospital of Philadelphia, Philadelphia, PA, USA

Abstract
Sickle cell disease (SCD) is a severe inherited hemoglobinopathy, challenging to study due to its low prevalence. Our study investigates the relationship between SCD and adverse pregnancy outcomes: preterm birth and Cesarean delivery (C-section). Of 50,616 women who gave birth at the Hospital of the University of Pennsylvania, 2,069 had SCD. SCD women were at decreased risk of preterm birth (8.7% vs. 9.27%) and increased risk of C-sections (36.78% vs. 34.5%) versus African Americans.

Introduction
Sickle cell disease (SCD) is a complex inherited genetic disorder and the most common hemoglobinopathy in the United States, affecting roughly 100,000 Americans1. SCD primarily affects individuals of African ancestry and is associated with high lifetime morbidity and premature mortality2. In addition, those pregnant with sickle cell disease have been shown to be at an increased risk of adverse outcomes3. Because SCD occurs in those of African ancestry it can be difficult to determine if the adverse pregnancy outcomes are due to SCD or because of health disparities common among the African American (AA) community4.

This study assesses the rate of pregnancy-related conditions and complications – specifically preterm birth and Cesarean delivery (C-section) – in AA patients, and specifically those with SCD.

Methods
We obtained data from the Hospital of the University of Pennsylvania (HUP) for 1,060,100 women treated at either in-patient or outpatient clinics between the years 2010 and 2017. We identified 50,616 women with delivery diagnoses and procedures. We then investigated two common pregnancy-related complications: preterm birth and C-sections.

To identify preterm birth diagnoses, and cesarean delivery procedures, we used relevant ICD-9 and ICD-10 codes. We identified our SCD population using diagnosis codes for SCD and found 2,069 females with SCD who gave birth at HUP. To investigate potential health disparities, we compared maternal outcomes between those with SCD, the AA population and also the general population of women giving birth at HUP. This study was approved by the Institutional Review Board of the University of Pennsylvania.

Results
At HUP, 50,616 unique women delivered and 2,069 of these also received a SCD diagnosis (4.09% of all deliveries at HUP). Overall, we found that 7.35% were preterm (3,709 patients), and 35.41% were Cesarean deliveries (17,925 C-sections). We found lower rates of preterm birth (8.7% vs. 9.27%) and higher C-section rates (36.78% vs. 34.5%) among patients diagnosed with SCD vs. the AA population. In addition, the AA population had higher preterm birth rates (9.27% vs. 7.35%), but lower C-section rates (34.5% vs. 35.41%) compared to HUP’s general population (Table 1).

Conclusions
In conclusion, studying pregnancy outcomes in those with SCD is complex due to a mixture of health disparities among African Americans and the disease state itself. Future work includes further distilling of the health disparities among the AA community and its role on SCD progression and adverse pregnancy outcomes.

References
Normalizing Medication Data in the UK Biobank
Michael N. Cantor, MD MA
Regeneron Genetics Center, Tarrytown, NY

Abstract
The UK Biobank, a significant resource for both clinical and genetic data, displays information about medications that participants were taking at the time of enrollment (field 20003). This list is not currently mapped to a standard, so contains entries that may be analyzed as individual traits. We used a set of reference terminologies to map medication entries to RxNorm CUIs. Of 6750 unique entries, we were able to map 91% using automated methods.

Introduction: The UK Biobank (UKB) (1) is a significant resource for both clinical and genetic data. Upon enrollment, participants respond to a number of questions, including self-reported medications (field 20003 in the UKB data showcase). This is the only data field that collects data on specific medications (rather than classes of medication), and data in this field is not standardized, so medications may be interpreted as individual, distinct traits, rather than being grouped together for analysis. For example, the list has separate entries for “Zantac” and “ranitidine”, and “Pulmicort” and “budesonide.” Mapping these medications to a reference ontology- in this case RxNorm- would provide more confidence in medication-related associations (2) and improve patient care through more extensive understanding of pharmacogenetic and gene-medication-disease interactions.

Methods: We obtained the medication list from the UKB data showcase, field 20003. We used the REST APIs and some form of string matching for all of the resources listed. Initially we used RxNAV to obtain RXCUIs (RxNorm Concept Unique Identifiers) from the raw, unedited entries (e.g. “timolol 0.25% eye drops) and then from edited entries with dosages and formulations removed. For the unmapped entries, we used the UK SNOMED CT drug extension as well as the UK’s dictionary of medicines and devices (DM+D), using those references to map to ingredients to obtain RXCUIs where the UK brand names were not available in RxNorm. For the remaining unmapped medications, we used the European Bioinformatics Institute’s (EBI) ontology lookup service; EBI’s chemical database, ChEMBL; and the Unified Medical Language System (UMLS).

Results: Though there are 6750 entries in field 20003, only 3674 are associated with at least one participant in the UK Biobank. However, for completeness we attempted to map the entire list. Using automated methods and these reference terminologies, we were able to map 91% of UKB medication entries to RXCUIs, and 93% of entries that were associated with at least one participant. The entries mapped to 1849 unique RXCUIs, and each unique RXCUI was represented between 1-65 times. Matching raw text and eliminating dosages and formulations produced 58% of the mappings; DM+D and SNOMED CT added 15%; and the remaining ontology tools produced the remaining 18% of mappings.

Discussion: The challenges in this project are typical of attempts to map self-reported data to a standard. Apart from the usual issues of mapping international brand names to a U.S.- based vocabulary (RxNorm), several of the brand-named medications that were represented in the DM+D dataset were not linked to their ingredients (e.g. product identifier that was not mapped to any active ingredients), possibly because the medication was no longer marketed, or if they were linked, they were only linked to their inactive ingredients. Additionally, though several medications without RXCUIs were initially mapped to the UMLS, these CUIs often did not map to specific ingredients or therapeutic categories, thus making these mappings unhelpful for categorizing specific drugs. Other issues included ordering of terms (e.g. “chlorothalidone / atenolol” vs. “atenolol / chlorothalidone”), ambiguous brand names, and unique misspellings that were not caught by approximate matching functions. Future work will address these issues programmatically and potentially use other vocabulary resources (e.g. OMOP’s ATHENA) to complete the mappings. Because of the idiosyncrasies of certain entries, we expect that at least some manual work will be required to achieve 100% mapping. The mappings will be available on request upon their completion and validation.

Conclusions: Standardizing medication names in the UK biobank is emblematic of the typical challenges of mapping self-reported data. Limiting input to only generic names would make standardization easier but would put an undue burden on participants; however, limiting input to a set of standard brand and generic names would most likely strike the correct balance.

References
Patient Timelines for Research-Oriented Exploration of Longitudinal Cancer Patient Data: PT Explorer

Rodrigo Carvajal, BS, Guillermo Gonzalez-Calderon, PhD, Miguel Betin-Montes, MS, Rhoda Gai-Cherry, MS, Jordan H. Creed, MPH, Travis Gerke, PhD, Joseph Markowitz, MD, PhD, Steven Eschrich, PhD, Anders Berglund, PhD

Moffitt Cancer Center, Tampa, FL

Clinical data integration has significantly advanced the ability to perform translational cancer research. Challenges still remain in the navigation of complex, high-dimensional and longitudinal patient data. Many systems provide a cross-sectional patient snapshot at a time point (e.g., current state) to enable efficient navigation. We apply clinical practice-centered navigation of complex patient timelines to create a framework for research-oriented data exploration. The goal is to provide a “movie” of the patient’s medical history as a dynamic timeline. Using the Moffitt Cancer Center (MCC) research data warehouse (111,667 research patients; 5,490 patient-oriented data elements), we developed the Patient Timeline Explorer (PT Explorer). We provide a multi-resolution, continuous view of the patient where patient encounters are classified and presented in a dynamic timeline. PT Explorer integrates data from clinic scheduling, billing, imaging, medication, biobanking, cancer treatments, and patient reported outcomes. The framework can be adapted to any encounter-based dataset with longitudinally captured events.

Currently available cohort exploration tools focus on drilling down to specific data elements, either for a single patient or for a cohort, then exporting these selected elements. PT Explorer allows multi-resolution access to a longitudinal series of data elements available from MCC’s data warehouse. Our case study involves a recent MCC data request for all data elements in a patient cohort. The current cohort explorer does not support this query and resulted in multiple linked spreadsheets. PT Explorer, in contrast, presents visual summaries of the data allowing for access to all variables for all patients within a cohort (see Figure 1), setting the stage for efficient downstream analysis planning, and making visualization more intuitive. Once a patient is selected, categories of data elements (low resolution) are represented as an integrated timeline (Figure 2). Specific longitudinal data types are represented compactly when drilling further into the data, for instance as heatmaps for symptoms assessment surveys (Figure 3). The views can be exported to different data formats (JSON, CSV, Excel), visualized as images, or accessible directly from analysis tools such as R.

With the availability and integration of clinical data elements in research data warehouses, the display and navigation of this information becomes crucial to the research enterprise. PT Explorer presents a “movie” of the patient clinical history from first physician visit in a web-based interactive timeline. This allows event-related data exploration and visual integration of timelines and, ultimately, better understanding of the cancer patient experience.
Addressing physical activity barriers among prostate cancer survivors through a peer-based digital walking program

Regina A. Casanova Perez MS,† Harsh Patel BS,† Savitha Sangameswaran MS†, David J. Cronkite MS,‡ Courtney Segal BA†, Dori Rosenberg PhD MPH,§ John L. Gore MD MS,¶ Jonathan L. Wright MD MS,¶ Andrea L. Hartzler PhD†

†Biomedical Informatics and Medical Education, University of Washington, Seattle, WA; ‡Kaiser Permanente Washington Health Research Institute, Seattle, WA; §Health Services, University of Washington, Seattle, WA; ¶Urology, University of Washington, Seattle, WA

Introduction
Prostate cancer (PC) survivors face significant barriers to survivorship-enhancing physical activity (PA). Although regular PA improves vitality, delays PC progression, and reduces the risk of PC mortality, only 30% of PC survivors meet national PA recommendations. Barriers to traditional structured exercise programs include lack of motivation, support, and exercise partners, which limit engagement and program impact. Although technology-based programs may improve access and reach, most lack social features many cancer survivors desire. Social media-based programs improve PA in young adult cancer survivors. Adding Fitbit tracking to such programs shows further promise. Yet, this approach has not been explored with older PC survivors. We know little about their design preferences or the efficacy of digital features to tailor such an approach. We engaged PC survivors in 3 human-centered design phases to develop and evaluate a digital program that combines technology and peer support to promote walking.

Phase 1. Assessing needs for improving PA: Using a series of 3 focus groups with PC support group attendees (N=61), we gathered needs and design preferences for our digital program to promote PA through personalized walking goals, supportive content, and peer support while facilitating step count tracking. Although most participants reported high interest in improving their PA level, there was agreement that “connections are what matters most to a group like us, not technology in isolation.” Preferred design features included: (1) high-quality content, (2) personal feedback on individual goals, and (3) support from small walking teams and peer mentors.

Phase 2. Designing a peer-based digital walking program: Informed by Phase 1, we built a program that includes: a self-selected ‘walking buddy’, a private Facebook group to access PA content and interact in group discussions, and a Fitbit Zip to track progress toward personalized walking goals. We developed a Python script to retrieve Fitbit steps.

Phase 3. Evaluating program efficacy for promoting PA over 6 weeks: We collected baseline step count for 1 week then held a kick-off meeting for participants (N=18) to meet, practice using the technology, and receive personalized walking goals to increase their average daily step count from baseline by 250 steps each week. Program evaluation followed a quasi-experimental pre-post design. After 6 weeks of program use, we found a significant improvement in average daily step count (mean increase of 2,463 steps/day, p<0.001). Exit surveys demonstrated moderate program acceptability with a mean(sd) System Usability Scale of 74(21). Exit interviews added qualitative insights into perceived program benefits, limitations, and future improvements. We are further investigating program engagement by analyzing week-by-week interactions with the Fitbit and Facebook group.

Conclusion
Findings establish a tested innovation, built in partnership with PC survivors, that demonstrates the potential to address PC survivors’ PA barriers in their everyday lives with a peer-based digital walking program. In future work, we plan to make improvements based on participants’ feedback and evaluate the program in a larger trial.

Acknowledgments: Fred Hutch/UW Cancer Consortium FHCRC 667963 and NIH/NLM T15LM007442

References
Usability Evaluation of a Clinical Decision Support Tool for Management of Peripheral Artery Disease Patients


Mayo Clinic Rochester, Minnesota.

Introduction: Practice guidelines for peripheral artery disease (PAD) patients recommend use of secondary prevention therapy with antiplatelet agents, statins, angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers, and smoking abstention. However, these strategies are underused by patients with PAD, with one study finding that only 16% of patients received all four recommended strategies in the community.1,6 Our study aimed to conduct a usability test on a clinical decision support (CDS) tool for populations of patients with PAD, entitled PAD cohort knowledge delivery solution (CKS), developed with the goal of assessment of guideline-based care received by PAD patients for quality improvement projects. The CKS tool links patient data with electronic knowledge bases to provide information/recommendations to providers at the point of care.5

Methods: We assessed both time savings and cognitive burden from the CKS compared to the electronic health record (EHR). The EHR was used to compare to the CKS because in the clinical workflow providers use information obtained by manual review of EHR for management of PAD patients. A set of 21 tasks were developed by study investigators to evaluate usability of the PAD CKS tool and the EHR system. Morae software was used to run the usability tests and to analyze data collected from these tests.2 Each scenario was also followed by an evaluation question regarding ease of completion of tasks within usability scenarios. The median for number of clicks, time taken, and number of keystrokes was calculated due to the skewed distribution of the data. Questions from the Quantitative System Usability Scoring System were included at the end of only the CKS usability test scenarios to generate a system usability score (SUS).1 Usability tests were conducted on 5 providers.4

Results: Mean system usability score (SUS) for the CKS was 90.5, with score > 80.3 considered excellent with a grade A rating. The median total amount of time taken to complete all 21 tasks in the EHR was 15.5 minutes compared to 4.5 minutes for all 21 tasks completed in the CKS (Table 1). EHR tasks had a median of 118 clicks in comparison to 31 clicks for the CKS, resulting in a difference of 98 clicks between the EHR and CKS. The EHR had a median of 26 keystrokes to complete the tasks, while the CKS had a median of 1 keystroke to complete tasks.

Table 1: Time and cognitive load for tasks with EHR and CKS

<table>
<thead>
<tr>
<th></th>
<th>Time Taken (min)</th>
<th>Number of Clicks</th>
<th>Number of Key-strokes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHR</td>
<td>15.5 (8.2, 30.5)</td>
<td>118 (67, 479)</td>
<td>26 (6, 64)</td>
</tr>
<tr>
<td>CKS</td>
<td>4.5 (2.7, 5.9)</td>
<td>31 (11, 43)</td>
<td>1 (1, 13)</td>
</tr>
<tr>
<td>EHR-CKS</td>
<td>9.8 (5.2, 27.8)</td>
<td>98 (36, 441)</td>
<td>22 (-7, 63)</td>
</tr>
</tbody>
</table>

Table 1 Legend: Data summarized as median (min, max) across the 21 tasks based on 5 providers who used CKS and 4 who used the EHR for usability tests. Difference based on data from 4 providers who used both systems.

Conclusions: Usability evaluation demonstrated that the PAD CKS tool saves provider time while using fewer clicks and key-strokes compared to EHR system for assessment of use of guideline recommended strategies by PAD patients. This tool may be used for quality improvement projects for patients with PAD.

References
EHR Workflow Integration Using Watson for Oncology in Colorectal Cancer:
Time-effectiveness and Information-comprehensiveness

Chun-You Chen, MD1,2,3, Yu-Chuan Li, MD, PhD3,4, Anita M. Preininger, PhD5, Yull Edwin Arriaga, MD5, Suwei Wang, PhD5, Metasebya Solomon, PhD5, Chintamani N. Bagwe, PhD5, Gretchen Purcell Jackson, MD, PhD5, Syed Abdul Shabbir, MD, PhD3,4

1Taipei Medical University (TMU), Taipei, Taiwan: 2Taipei Municipal Wanfang Hospital, TMU, Cancer Center; 3Graduate Institute of Biomedical Informatics, TMU; 4International Center for Health Information Technology (ICHIT), TMU; 5IBM Watson Health

Introduction
An artificial intelligence clinical decision-support system (CDSS), Watson for Oncology (WFO), processes structured and unstructured patient data using natural language processing and machine learning. WFO identifies therapeutic options for cancer patients based on patient-specific information, called attributes. Examples of attributes abstracted by WFO include age, gender, stage, metastatic disease, imaging measurements, labs, prior treatments and co-morbidities. WFO provides a set of evidence-informed treatment options based on training cases provided by MSK cancer experts. Attributes can be entered in WFO manually or automatically by integration with an EHR system. An EHR synced tool like WFO provides value for having to incur the burden of data entry into an EHR. This is a retrospective study comparing EHR-WFO integration versus manual WFO data entry in colorectal cancer patients.

Methods
WFO was implemented in 3 Taipei Medical University (TMU)-affiliated hospitals; only one hospital integrated WFO with their EHR system. The EHR-WFO integrated hospital automatically processed cases after the first report of cancer by a pathologist, while in the other two hospitals, attributes were entered in WFO manually. Patient age, gender, imaging and pathology reports identified on or after the index date (14 days prior to diagnosis) were integrated as part of the relevant EHR for each case. Colorectal cancer patient attributes entered manually versus those abstracted by the integrated EHR-WFO system from 1/2018 to 2/2019 were compared. Shapiro-Wilk’s test for normality of variables summarized the data as median. Significance of differences was determined by Wilcoxon sign-ranked test.

Results
Of the 140 cases analyzed by WFO, 63 (45%) used the integrated EHR-WFO system to abstract attributes; 77 cases (55%) required manual data entry into WFO. EHR integration significantly decreased manual attribute entry (Figure 1, median values 14 vs. 16, respectively, \(P = .00014\)). EHR-WFO integration abstracted significantly more attributes into WFO, including optional ones, than manual input alone (median 23 vs. 16, respectively, \(P < .0001\), Figure 2).

Conclusions: EHR-WFO integration significantly increased the number of attributes entered into the system for treatment decision making and decreased manual attribute input, potentially saving time and enhancing the comprehensiveness of data. These findings suggest that EHR-WFO integration increased the richness of data available for treatment decision making beyond what WFO minimally requires, as optional attributes may be omitted during manual entry. The types of attributes extracted by each method and the effects of additional attributes extracted on decision making are under further study.
Ensemble Imputation for Healthcare Data

David Chen, PhD1, Christopher G. Scott, MS2, Christina L. Luong, MD3, Itzhak Zachi Attia, MS3, Che G. Ngufor, PhD1, Adelaide M. Arruda-Olson, MD, PhD3, Patricia A. Pellikka MD3

1Division of Digital Health Sciences, 2Division of Biomedical Statistics and Informatics, 3Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN

Introduction
Missing data is a common problem afflicting the vast majority of healthcare data[1]. Imputation methods can bias the final model and reduce the predictive accuracy. In this work, we combine bootstrap aggregation and imputation to capture the uncertainty in the imputed missing values and reduce the variability in the final learned model[2]. We demonstrate bootstrap imputation (BI) in an analytic task of predicting 5-year mortality in patients who have undergone a stress echocardiography test.

Methods
Figure 1 compares the schematic of conventional single imputation (SI) with the proposed algorithm. No imputation, null, mean, hot deck, and multiple imputation methods were compared. A decision tree (C5.0) and gradient boosting machine (GBM) methods were used to model. Area under receiver operating curve (AUC) metric with 5-fold cross-validation was used to evaluate predictive accuracy. All models were trained and evaluated in R.

A retrospective dataset of 68,706 patients was created from all adult patients who underwent stress echocardiography exam at Mayo Clinical Rochester. A total of 150 variables were abstracted from echocardiographic data, with missingness ranging from 0% to 48% and median missingness of 4.1%.

Results and Discussion
Table 1 shows comparison of results of the various imputation methods. The mean improvement in AUC using BI over all imputation methods for was 0.029 for decision trees and 0.006 for GBM. More representative imputation methods (hot deck or multiple imputation) show improved performance compared to simple ones. We hypothesize that single imputation cannot capture the uncertainty of the imputed missing value. Therefore a poor imputation model can severely bias the final classifier. Second, the imputation itself maybe unstable. Using bootstrapped sampled data for imputation reduces overfitting due to the imputation.

Table 1: AUCs of various imputation methods. Single and bootstrap refers to single imputation vs aggregation imputed bootstrap sampled data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Decision tree</th>
<th>Gradient boosting machine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single</td>
<td>Bootstrap</td>
</tr>
<tr>
<td>No Imputation</td>
<td>0.641±0.015</td>
<td>0.674±0.018</td>
</tr>
<tr>
<td>Null</td>
<td>0.634±0.019</td>
<td>0.652±0.030</td>
</tr>
<tr>
<td>Mean</td>
<td>0.640±0.018</td>
<td>0.668±0.026</td>
</tr>
<tr>
<td>Hot Deck</td>
<td>0.641±0.016</td>
<td>0.690±0.017</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>0.685±0.019</td>
<td>0.714±0.024</td>
</tr>
</tbody>
</table>

Conclusion
The results presented in this work demonstrate how ensemble based imputation can lead to more robust classification for a complex healthcare analytics task. The presented method allows for flexibility depending on the characteristics of the dataset and the limitations of the computing environment.

References
1. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Egems 2013;1(3)
Bipartite Network Analysis with Multiple Datatypes: Implications for Precision Medicine in the Age of Multi-omics Data

Tianlong Chen PhD¹, Yu-Li Lin MS², I-Chia Liao MPH³, Yong-Fang Kuo PhD², Emmanuel Santillana MS⁴, Clark Andersen MS², Laurel A. Copeland PhD³, Suresh K. Bhavnani PhD¹

¹Inst. for Translational Sciences, ²Preventive Medicine and Community Health, Univ. of Texas Medical Branch, Galveston, TX; ³Center for Applied Health Research, Baylor Scott & White Health, Temple, TX

Abstract

Bipartite networks, which simultaneously model both patients and their characteristics, have been useful for identifying significant and comprehensible patient subgroups, a critical step in designing targeted interventions. However, because clustering algorithms require uniform data ranges to identify subgroups, bipartite networks have been used mainly to model uniform data types such as only the binary status on comorbidities, or only the continuous range-normalized values of gene expressions. Here we demonstrate the use of predictive probabilities to transform variables of multiple datatypes into a uniform range, which enabled the use of bipartite networks to identify and comprehend significant subgroups of readmitted patients.

Introduction

Because bipartite networks model both patients and their characteristics, they have been useful in not only identifying significant and replicated patient subgroups, but have also enabled clinicians to interpret why patients in each subgroup have negative outcomes, critical for designing interventions. However, because biclustering algorithms such as modularity require uniform ranges to enable the partitioning of patients and characteristics into subgroups, they are limited to modeling variables that have uniform ranges such as only the binary status of comorbidities, or only range-normalized gene expressions. Here we use predictive probabilities to transform multiple datatypes into a uniform range, enabling us to simultaneously model all relevant clinical variables like number of opioids (continuous), marital status (categorical), and presence of anemia (binary).

Method and Results

Data: We extracted 100% cases (total hip arthroplasty [THA] patients readmitted within 30 days of hospital discharge) and controls (not readmitted within 90 days, matched by gender, age and race) from the 2006-2015 data in the Veteran’s Administration Clinical Data Warehouse (VA-CDW). This resulted in 460 matched pairs each in the training dataset and in the replication dataset. Variables consisted of those identified by a physician and occupational therapist that were high risk for readmission, and all comorbidities from the Charlson, and CMS-CC indices that were also in the VA-CDW, resulting in 47 variables (41 binary, 1 categorical, 5 continuous). As domain experts prefer up to 10 variables when interpreting patient subgroups, we used logistic regression to estimate the significance of each variable for predicting readmission in each dataset, and selected the top-10 in the training dataset that were also highly ranked (in the top-20) in the replication dataset. Next, we transformed the value of each variable for each patient into the respective univariable probability (range 0-1) for predicting readmission. As we used a 1:1 case-control study design, the readmission probability in the study cohort is 0.5. We therefore dichotomized all the probabilities (<=.5=0; >.5=1) to conduct the bipartite network analysis.

Method: The analysis consisted of two steps. (1) Identification of patient subgroups by representing the data as a bipartite network where nodes represented either patients or variables, and the edges represented the dichotomized predictive probability of a variable. The training network was analyzed using bicluster modularity to calculate the number and boundaries of the biclusters (containing patients and variables), and the degree of biclustering (Q). The significance of Q was calculated by comparing it to a distribution of Q generated from 1000 random permutations of the network. The resulting patient subgroups were visualized using the ExploreLayout¹ algorithm. (2) Replication testing by repeating the above analysis on the replication dataset, and using the Rand Index (RI) to measure the proportion of pairs of comorbidities that co-occurred and did not co-occur within the same cluster in both networks. RI significance was calculated by comparing it to a distribution of RI generated from 1000 random permutations of both networks.

Results: As shown in Fig. 1, the analysis identified 7 biclusters with significant modularity (Q=0.35, p<0.05), and significant replication (RI=0.87, p<0.05). The visual separation of the biclusters in the network layout enabled the clinicians to infer the reason for readmission in each bicluster. For example, the patient subgroup in the upper right hand corner had a combination of gait disorder and peripheral vascular disease requiring different interventions compared to the patients with COPD. The results suggest that the method could be useful in analyzing multi-omics data to identify patient subgroups for precision medicine applications.

Acknowledgements. Funded in part by a pilot grant from the HCSR-N-OAICs AGING Initiative.

References


Fig. 1. Seven biclusters of readmitted THA patients.
Review and Analysis of Current Mobile Apps for Depression
Lewis Chesebrough, M.S.¹, Annie Myers, M.A.¹ Lingchen Lou¹, Ruixuan Hu¹, Meghan Reading Turchioe, PhD, MPH, RN¹, Ruth Masterson Creber, PhD, MSc, RN¹

¹Weill Cornell Medicine, Department of Healthcare Policy and Research, New York, NY

Introduction
Clinical depression, defined as having one major depressive episode in a given year, affects 8.1% of American adults age 20 and over, and 37% of adults in all ages received no treatment. (1) Although psychological and pharmacological treatments for depression exist, many patients do not receive formal care and remain undiagnosed. (1) Consumer mobile apps have potential to improve patient outcomes and increase access to care but a systematic review of the quality and functionality of apps for depression has not been completed.

Objective
Using a standardized methodology, examine the quality of mobile apps targeting individuals with depression.

Methods
We completed a systematic review of mobile apps for depression self-management by entering six relevant search terms in the Apple, Android, and Amazon app marketplaces: "depression", "mood improvement", "mental health self-management", "depression monitoring", "depression screening", and "depression self-management" We selected apps using four rounds of ad hoc inclusion and exclusion criteria. We downloaded and independently reviewed the final apps and rated them according to the Mobile Application Rating Scale (MARS)(2), IMS Institute for Healthcare Informatics functionality scores (3), and general guidelines for depression management in the absence of patient-specific self-management guidelines. (5) We also conducted a literature review to assess whether the selected apps had been tested in clinical trials.

Results
Of 1198 potentially relevant apps searched, 31 apps met inclusion criteria. Most apps were excluded because they were unrelated to depression, were designed for educational purpose and diagnoses only, or were games. Inter-rater reliability between reviewers was high (0.75 – 0.83). Pacifica had the highest average MARS total score (4.6 out of 5). More than half of the apps (23/31, 74%) had acceptable MARS scores (≥3.0) and among those, 10 apps had MARS scores ≥ 4.0. Pacifica also had highest score for engagement (4.8), satisfaction (4.6) and behavior change (4.8). Elevat (4.4) had the highest score for behavior change. For functional score, the top three apps were Snapclarity, Pacifica, and Joyable. For Depression Management Features, the top performing apps were AkiliHealth, Booster Buddy, and InnerHour. Only one app, Woebot, was tested in an RCT, which showed that compared to a control group provided information on depression self-management, the app intervention group had significantly reduced depression symptoms. (5)

Conclusion
Mobile apps for depression exhibit a wide range of quality, but more than half of the apps included in this review had acceptable quality. The MARS scores correlated with IMS Functionality Scores, but not with the number of features for depression self-management. The high scoring apps in this review show that mobile health interventions can increase access to support, help manage symptoms of depression, and improve quality of life.

References
Reading Emotional Support Messages on a mHealth App Predicts the Initial Lapse after Alcohol Detox

Ming-Yuan Chih, PhD1, Woohyun Yoo, PhD2, Dhavan V. Shah, PhD3, Andrew Quanbeck, PhD3, David H. Gustafson, PhD3

1University of Kentucky, Lexington, KY; 2Incheon National University, Incheon, South Korea; 3University of Wisconsin-Madison, Madison, WI

Introduction
A key feature of Addiction-Comprehensive Health Enhancement Support System (A-CHESS),1 a smartphone app for relapse prevention after alcohol detox, is to provide preemptive services (e.g., sending alerts to addiction counselors) when patients are likely to lapse based on a built-in risk assessment/prediction model2. However, the predictability of this model can be improved.3 Based on the theories of emotional support, a computer-aided content analysis was conducted to analyze messages in an online forum provided in A-CHESS.4 These online peer-to-peer messages filled with emotional support contents may influence the lapse risk of patients who read them. We reported here the results of an effort to integrate these emotional support message coding data into the prediction model. The hypothesis is that reading messages with emotional support contents may affect patient’s imminent risk to lapse.

Methods
Data were from the intervention arm of a randomized control trial testing A-CHESS.1 The dataset—142 patients (51 initial lapses and 91 matched non-lapses)—in a prior study was used.3 2745 messages from the online forum in A-CHESS were retrieved. A hybrid human-computer coding method, reported elsewhere4, included the following steps: (1) developing theory-based coding constructs; (2) creating dictionaries of keywords in each construct; (3) establishing relationships between multiple terms/concepts in two software programs, WordStat 6.1 and QDA Miner 4.1; and (4) running the computer programs to generate coding scores in each coding construct. These message coding data were merged with the log file-derived messages reading activities 7 days prior to the prediction timepoint.3 A reading event was defined as the action when a user clicked on the link to open a particular message. Blocked logistic regression model was used. Three blocks included gender (1st block), 10 items in the weekly self-reported recovery status (e.g., urge) (2nd block), and the total scores of each of the 9 emotional support message coding constructs of the messages read 7 days prior (3rd block). Between-block model comparison was done by comparing changes of the chi-square and the degree of freedom values. Area Under receiver operating characteristic Curve (AUC) analysis was reported.

Results
Every new block accounted for significant model variance (see Table 1). Therefore, adding message coding improved the prediction model. On average, the most read coding construct in a week is Encouragement (mean=4.27) and then Care (mean=1.97). Based on the full model, confidence staying sober (p=0.02) and reading more messages containing Prayer words (such as “God with you”, p=0.044) were significantly associated with a reduced risk to lapse. Interestingly, reading more messages with Empathy words (e.g., “sorry” or “worry”) and Encouragement words (e.g., “cheer” or “hope”) was associated with an increasing risk to lapse in trending p-values (0.098 and 0.079 respectively). AUC is 0.862 with 95% CI [0.8, 0.92].

Table 1: Logistic Regression Analysis

<table>
<thead>
<tr>
<th>Block</th>
<th>Chi-Square</th>
<th>df</th>
<th>ΔChi-Square</th>
<th>ΔDf</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>5.392</td>
<td>1</td>
<td>---</td>
<td>---</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Gender+Recovery</td>
<td>36.119</td>
<td>11</td>
<td>30.727</td>
<td>10</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Gender+Recovery+</td>
<td>61.698</td>
<td>20</td>
<td>25.579</td>
<td>9</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Conclusion
Adding reading emotional support related contents in an online forum can improve the predictability of the initial alcohol lapse after detox. This study was supported by grants (R01 AA017192 and R01 AA24150) from NIAAA.

References
Perceptions of the Use of a Real-time Medication Monitoring Pill Bottle linked to a HIV Self-management App

Hwayoung Cho, PhD, RN1, Gabriella Flynn, MED2, Maureen Saylor, MPHc2, Melissa Gradilla, MPH2, Rebecca Schnall, PhD, MPH, RN, FAAN2

1College of Nursing, University of Florida, Gainesville, FL, USA; 2School of Nursing, Columbia University, New York, NY, USA

Introduction
Non-adherence to antiretroviral therapy (ART) among persons living with HIV (PLWH) remains a critical problem. Mobile health (mHealth) technology has the potential to be a useful tool in self-management of HIV to improve health outcomes such as medication adherence, engagement in healthy behaviors, and health-related quality of life.1-5 To improve long-term medication adherence using effective interventions that enhance the HIV care continuum, the ‘Wise App’, an HIV self-management app with real-time medication monitoring, was developed (Figure 1). The purpose of this work was to conduct an in-depth analysis of factors influencing use of the electronic pill bottle linked to the Wise App among PLWH.

Methods
This study was conducted as part of a larger randomized controlled trial with 200 PLWH who were assigned to the intervention or control group (1:1) through block randomization. In-depth interviews were conducted using a semi-structured interview guide at the 3-month follow-up visit with intervention group participants. All interviews were audio-recorded and transcribed. We explored emerging themes with similar patterns across interviews through thematic analysis and organized the themes according to the constructs of the Fit between Individuals, Task and Technology (FITT) framework.6

Results
Thirty-eight PLWH participated in the interviews. Nine themes organized by the FITT framework were identified. Three themes related to the fit between individuals and task were: motivation for strict medication adherence, self-efficacy for overall health management and engagement with medication reminders. Four themes related to the fit between individual and technology were: ease of use, HIV-related stigma and disclosure of HIV status, customized alert of medication time windows based on individual routine set-up and preference for device design. Two themes related to the fit between task and technology were: system functionality of data transfer from the electronic pill bottle to the app and self-awareness of system syncing signals.

Conclusions
This work demonstrated that tracking medication adherence and receiving push-notification medication reminders through the electronic pill bottle connected to the app supports PLWH in adhering to their medication regimens. Findings from this study highlight the importance of adequate consideration of the needs of intended users in designing customizable mobile health technology, including HIV-related stigma, disclosure of HIV status and ART regimens.

Acknowledgements This study was supported by the Agency for Healthcare Research and Quality under award number R01HS025071 (PI: Schnall).

Reference
Changes in Nursing Activity After Implementing a CDS Service Predicting the Risk of Falling Based on Electronic Medical Records Data

Insook Cho, PhD1, Insun Jin, PhD2
1Inha University, Incheon, Republic of Korea; 2Public Health Insurance Ilsan Hospital, Goyangsi, Republic of Korea

Abstract

One-third of inpatient falls are preventable, but preventing them is hindered by the inability to accurately estimate their risk. A well-designed clinical decision support (CDS) service can help nurses decide how to allocate their services and resources. Here we report preliminary prospective observations obtained from the implementation of a CDS service designed to predict the risk of falling for individual patients based on electronic medical records data captured during nursing care.

Introduction

The quality of hospital care would be greatly improved if nurses are able to predict patient events that are likely to occur within 24 hours based on data available in electronic medical records (EMR). We developed a conceptual predictive model for estimating the risk of falling at the individual level using data captured routinely by an EMR system. The model was a Bayesian belief network with 48 nodes, 80 links, and 11 million conditional probabilities, and it was cross validated at 2 hospitals with different EMR systems and different nursing vocabularies. We used the model to implement a clinical decision support (CDS) service that predicts the daily probabilities from EMR data at a metropolitan hospital in Seoul, Republic of Korea.

Methods and Results

This prospective cluster randomized trial involved 12 nursing units of medical-surgical, neurology, and geriatrics departments. It was implemented in May 2017 and designed to finish in April 2019. The 12 units were randomly allocated to the experimental and control groups. The experimental units provided the CDS service and the control units provided their usual care. According to the data analysis frame defined in our previous work, we monitored the daily frequencies of nursing assessment and intervention activities provided to patients during two time periods: the first evaluation (6–12 months after implementation), and the second evaluation (12–18 months after implementation). We used the existing information as baseline (January 2015 to March 2017) for the experimental group. The assessment activities increased about 54-fold in the experimental group compared to the baseline, but the interventions had not increased after 12 months and only slightly after 18 months (Table 1). The assessment activities in the control group remained at similar levels throughout, but there were more interventions than in the experimental group after 12 months.

Table 1. Comparison of daily frequencies of fall risk assessments and interventions during at-risk days

<table>
<thead>
<tr>
<th>Activity</th>
<th>Time</th>
<th>Baseline</th>
<th>1st evaluation (6–12 months)</th>
<th>2nd evaluation (12–18 months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assesment</td>
<td>Exp.</td>
<td>1.05 (1.03–1.07)</td>
<td>54.49 (54.22–54.76)</td>
<td>57.81 (57.63–58.00)</td>
</tr>
<tr>
<td></td>
<td>Cont.</td>
<td>NA</td>
<td>12.82 (12.64–13.00)</td>
<td>13.91 (13.74–14.08)</td>
</tr>
<tr>
<td>Intervention</td>
<td>Exp.</td>
<td>20.38 (20.26–20.50)</td>
<td>20.87 (20.54–21.20)</td>
<td>29.21 (28.81–29.61)</td>
</tr>
<tr>
<td></td>
<td>Cont.</td>
<td>NA</td>
<td>33.71 (33.12–34.30)</td>
<td>35.12 (34.78–35.61)</td>
</tr>
</tbody>
</table>

Exp., experimental; Cont., control; NA, not applicable.

Data are mean (95% confidence interval) values.

Conclusion

These preliminary results give a clue that the CDS service predicting the daily risk of falling based on EMR data may affect positive changes of relevant nursing activities in terms of efficiency and resource allocation. To confirm it, further data collection and detailed analyses of nursing processes depending on patient risks are needed.

Reference


Acknowledgement: This study was supported by grants of the Korea Healthcare Technology R&D Project, Ministry for Health, Welfare & Family Affairs, Republic of Korea (No. HI17C0807) and National Research Foundation of Korea (No. NRF-2019R1A2C2007583)
Effective Use of Secure Text Paging for Subspecialty Consult Requests

Chi D. Chu, MD1, Delphine S. Tuot, MDCM MAS1, Jonathan Duong, MD2, James D. Harrison, PhD2, Raman R. Khanna MD MAS2

1Division of Nephrology, University of California San Francisco, San Francisco, CA; 2Division of Hospital Medicine, University of California San Francisco, San Francisco, CA

Introduction
Despite the widespread prevalence of electronic health records, subspecialists receiving consult requests often benefit from an informational context provided via phone or in-person conversation with requesting clinicians. Based on previously developed frameworks for teaching effective consultation, the “best practices” for providing such context include a description of patient identifiers, the reason for consult, the level of urgency, and caller contact information. Alphanumeric text paging to request subspecialty consultation allows some context to be provided in the content of the page itself; however, the most effective use of this limited medium for providing such context is not known.

Methods
We extracted 12,290 secure text pages sent at the University of California, San Francisco Medical Center between 4/5/2017 and 11/17/2018 containing the case-insensitive string “consult” or “c/s”. From this, we randomly sampled 30 pages requesting consults sent to each of the 7 most consulted medical subspecialties, for a total of 210 text pages. Messages were coded independently by three raters (C.C., D.T., R.K.) for inclusion of elements based on literature review on consultation “best practices”: caller details (name, service, phone number, pager number), reason for consult, and indication of urgency. Additionally, we assigned a global quality rating (5 categories from “Very Poor” to “Very Good”). We used multivariable ordinal logistic regression to determine predictors of higher quality pages.

Results
The proportion of pages that included each best practice element is summarized in Figure 1. Of these elements, most pages included basic contact information and reason for consult; fewer than 20% explicitly indicated urgency. Of these 6 elements, the majority of pages included at least 4 or more elements (Figure 2).

Figure 1: Elements Included in Consult Pages

Figure 2: Consult Page Completeness

In multivariable ordinal logistic regression, the most important element associated with higher page quality was reason for consult (OR 22.35; 95% CI 8.10 – 61.66, p < 0.01), followed by callback number (OR 6.19; 95% CI 0.77 – 49.50, p = 0.09), caller name (OR 5.04; 95% CI 1.93 – 13.12, p < 0.01), and urgency indication (OR 3.28; 95% CI 1.59 – 6.72, p < 0.01).

Conclusion
Despite limitations of the medium, the majority of text paged consult requests contained 4 or more informational elements; however, urgency was infrequently included despite being traditionally included in “best practices” for consultation. This suggests there may be differences in elements influencing effectiveness of consults requested by text page as opposed a telephone conversation; thus, established best practices for telephone consultation may not apply directly to text paged consult requests: this is a potential area for future research. Our results suggest that clinicians requesting consults via text page should be aware that reason for consult was by far the most important element.
Hospital Performance on Unsafe Medication Orders in the 2017-2018 CPOE Evaluation Tool

Zoe Co1, Diane Seger1, RPh2, Lisa Newmark, BA3, Melissa Danforth, BA3, David C. Classen, MD, MSc4, David W. Bates, MD, MSc1,2,5

1Department of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, 2Clinical and Quality Analysis, Partners Healthcare, Somerville, MA, 3The Leapfrog Group, Washington D.C., 4The University of Utah, Salt Lake City, UT, 5Harvard Medical School, Boston, MA

Introduction: Although electronic health record (EHR) and computerized physician order entry (CPOE) implementation is now widespread, evidence suggests that these systems do not always result in reduced harm.1 Every year, hospitals in the U.S. take the CPOE Evaluation Tool via The Leapfrog Group’s annual Hospital Survey, to assess their EHR system’s ability to alert prescribers to common and serious prescribing errors, as implemented.2,3 The results of the tool provide hospitals with an overall percentage score of potentially unsafe orders detected, as well as a fatal orders analysis, which lists the potentially fatal orders that a hospital failed to stop or alert on. We evaluated how hospitals’ EHR systems performed overall and on potentially fatal orders in 2017 and 2018.

Methods: Detailed datasets containing the 2017 and 2018 CPOE Evaluation Tool results were extracted for analysis. The datasets only included the highest test score for each hospital, and any incomplete tests were removed. We also only included electronically orderable test orders by hospitals, meaning if a certain medication was not electronically orderable by a hospital, that test order was removed from their denominator during scoring.

To analyze how hospitals performed against potentially unsafe orders, we calculated the mean overall percentage score, and the mean overall fatal order score. For the 2017 and 2018 versions of the test, the fatal orders were distributed throughout the following order categories: drug dosing (both single and daily dosing), drug laboratory, drug route, and drug-drug interaction.

A secondary analysis was performed, in which we compared hospitals’ responses to the same fatal orders between the 2017 and 2018 test.

Results: From the 2017 data, we evaluated results from 1,750 hospitals. There was a total of 6,591 orderable fatal orders distributed throughout the test. Of these, 1,603 (24%) did not cause an alert to fire. Within these orders, drug dose daily was the fatal order category that hospitals performed the worse in, with 933 (58.2%) orders not alerted on. In contrast, drug-drug interaction was the category that hospitals performed the best in, with only 21 orders (2.25%) not alerted on.

The mean overall score in 2017 was 58.0%, while the mean fatal orders score was 76.1%. More than half (55.4%) of the hospitals received a fatal order score of 100%, while 44.6% of hospitals got at least one fatal order incorrect. Of these hospitals, 113 hospitals did not alert on any fatal orders.

Comparing these results to 2018, hospitals showed improvement. The mean overall score in 2018 was 65.8%, while the mean fatal order score was 83.5%. There was a total of 5,486 fatal orders, and 933 fatal orders (17.0%) were not alerted on. The percentage of hospitals that got at least one fatal order incorrect decreased to 35%. Drug dose daily was still an area that hospitals struggled with. However, the percentage of daily dose orders not alerted on decreased to 43.4% (405 orders).

From the secondary analysis, we identified 422 hospitals that received the same fatal order in 2017 and 2018. Of these hospitals, 81.3% (343 hospitals) got the same orders correct in both years, and 11.1% (47 hospitals) alerted on the fatal order they did not alert on in 2017. However, 7.6% (32 hospitals) got the same order wrong in both years.

Conclusion: Although the mean overall score suggests that most hospitals’ CPOE systems alert on most common prescriber errors, there is still important room for improvement in how systems respond to potentially fatal orders. These results reveal critical areas of improvement that hospitals can use to improve their EHR system. The tool is being expanded to include other domains, which will hopefully help hospitals improve their safety further.

References:
Impact of EHR Usability on Provider Efficiency and Patient Safety in Non-Hospital Settings

Jamie Cole, P.S.M. ¹, Wanda Sankey B.S.¹, Guenter Tusch, Ph.D¹, Raymond Higbea, Ph.D., FACHE¹, Marie Vanderkooi, DNP, MSN, RN-BC¹, Larry Warkoczeski, M.H.S.A. J.D.¹

¹Grand Valley State University, Allendale, MI, USA

Abstract

Electronic health records (EHRs) have been implemented in many healthcare settings under the premise of increased productivity and improved patient interactions. However, severe unintended consequences from implementation and design of these systems have emerged. Poorly implemented EHR systems may endanger the integrity of clinical or administrative data. That, in turn, can lead to errors that may jeopardize patient safety or decrease quality of care. Therefore, analysis of the benefits anddownfalls of the implementation of EHRs is paramount step to improving provider-patient interactions and overall public health. Prior review of the implementation of these record keeping systems has been primarily focused on the hospital setting. Given the variability and importance of non-hospital medical centers, assessing the impact of EHRs to local offices is of exceptional importance. This study aims to better assess the implementation of EHRs by electronic survey and group interviews of local medical centers in an effort to improve patient safety and EHR usability.

Introduction

EHR usability refers to the efficiency and effectiveness of the system’s use and the satisfaction of the users when accomplishing specific tasks within their environment; this includes user–friendly workflow design supporting efficient, effective quality care. The lack of standard user interfaces remains a challenge for clinicians who work in multiple care settings potentially influencing provider productivity and patient safety. EHR-related errors have been documented in the past, however this documentation has been primarily focused on the hospital setting. Identifying the errors and benefits of implementing EHR in non-hospital settings is still needed to improve these patient-provider interactions. The purpose of this project is to identify the impact of EHRs on local medical offices as a means to improve local implementation of EHRs as well as identify common hindrances to the efficacy of EHR implementation in a non-hospital setting.

Methods

An initial literature review of current EHR studies was performed by reviewing over 300 articles from various databases including PubMed, ProQuest, Google Scholar. These articles included inpatient and ambulatory settings and were used as a basis to construct a series of interview questions. These questions focused on the categories and subcategories of provider centric (sub-categories: design, workload, implementation), patient centric (sub-categories: care coordination, patient safety, quality) and outcomes (sub-categories: patient-provider relationship, health outcomes). Through a partnership with the Michigan Medical Group Management Association, local medical offices were interviewed as well as surveyed on these topics. Finally, their responses were recorded and analyzed in an effort to identify shortcomings and benefits EHRs have in local medical offices.

Results and Conclusion

In this study, a series of questions were developed based on a current literature search of EHR implementation. The focus of these questions were established during a previous phase of this study, where a literature review of EHR impact demonstrated three major categories (provider centric, patient centric, and outcomes) and eight sub-categories (design, implementation, workload, safety, quality, care coordination, patient-provider relationship, and health outcomes) were heavily impacted by EHRs. Based on this literature review, a model was established and the effectiveness of the implementation of EHRs was analyzed through the lens of this model. This study is ongoing and the data collection and analysis continue to grow. It is anticipated that this study will be completed over the next 1-1.5 years. After assessing the advantages and disadvantages of EHRs, this information will be shared with these medical offices as well as possibly establishing coordination between the medical offices and Grand Valley State University to assist in the improved implementation of EHRs to office usability and patient safety.

1554
Automated Prediction of Patient Stability and Discharge within 36 Hours

Kevin Coppa,1 Marsha Meytlis, PhD,1 Jamie S. Hirsch, MD,1,2 Jan Horsky, PhD3
1Department of Information Services, Northwell Health, Lake Success, NY
2Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY
3Center for Research Informatics and Innovation, Northwell Health, Manhasset, NY

Introduction

Extensions of hospital stays that do not add clinical value over necessary treatment and stabilization are associated with increased risk of iatrogenic infection, poor nutrition and unwarranted higher care utilization.1 Planning of care in the community and completion of documentation usually starts after a patient discharge decision has been made, often adding hours and days to stay. Advance identification of stable patients who will be soon ready for discharge would allow starting the administrative process early, reducing exstraneous hospitalization days. Data routinely collected and stored in the EHR can be automatically analyzed by machine learning algorithms to calculate how likely a specific patient is to be stable enough for discharge within 24 to 36 hours. This pilot study compared four classification algorithms and a regression model for best performance in predicting patient stability within 36 hours: a two-class support vector machine, a random forest model, decision tree, a neural network and a logistic regression.

Methods

Algorithms were trained on 90% of available data and their predictive ability tested on the remaining 10%. A set of suitable clinical markers and salient patient characteristics that a hospitalist would likely consider when making a discharge decision was derived from discussions with clinical experts. It included age, data flows, order-related data and recent laboratory test results that were entered as features into the models to approximate a decision-making process. Binary values represented the presence and absence of orders for oxygen, pain management, telemetry and imaging studies, antineoplastic, steroid, and diuretic medications, the need for blood transfusions and the presence of an active diet order. Parameters related to the stay were the number of orders and notes in the preceding 24 hours, weekday and inpatient unit. Laboratory results included oxygen saturation, temperature, hemoglobin, white blood cell count and sodium level and were converted to binary markers as normal or abnormal. The most recent LACE score, Charlson Comorbidity Index and MEWS score were also included. We queried a clinical data repository that contained 76,587 visit days from 14,804 unique patient stays. Eliminating rows with incomplete data reduced the set to 74,129 visit-days from 13,864 unique admissions. Performance of the algorithms was then evaluated on this dataset.

Results

Non-customized random forest model using optimized hyper parameters showed better performance than its competitors (Table 1), with positive predictive value of 48% and a balanced accuracy of 70.6% (95% CI: 0.690.71). (Figure 1). The four algorithms and logistic regression model had similar performance characteristics, and all performed better than the decision tree.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Sens %Spec</th>
<th>%Pos PV</th>
<th>%Neg PV</th>
<th>%Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log regression</td>
<td>73.5</td>
<td>64.5</td>
<td>46.8</td>
<td>85.1</td>
</tr>
<tr>
<td>Neural net</td>
<td>73.9</td>
<td>64.1</td>
<td>46.7</td>
<td>85.2</td>
</tr>
<tr>
<td>Decision tree</td>
<td>71.9</td>
<td>60</td>
<td>43.3</td>
<td>83.4</td>
</tr>
<tr>
<td>Sup vector mod</td>
<td>74.5</td>
<td>64.6</td>
<td>47.2</td>
<td>85.6</td>
</tr>
<tr>
<td>Random forest</td>
<td>76</td>
<td>65.1</td>
<td>48.2</td>
<td>86.4</td>
</tr>
</tbody>
</table>

Table 1 Algorithms predicting 36-hour patient stability

Figure 1 ROC, random forest

Conclusion

The random forest model using data routinely available in medical records provided guidance for the optimization of the discharge process, possibly avoiding unnecessary extensions of stay. This model likely performs better as it has intuitive decision rules, allows non-linear variables and accounts for their interaction. The usual length of stay during the study period was shorter than 2 days which likely increased the negative predictive values of all tested algorithms. The algorithm is currently used to prioritize the scheduling of imaging tests needed before discharge.

References

Mount Sinai Spanish-Language OpenNotes Navigator Program

Catherine K. Craven, PhD, MA, MLS1,2, Lina Jandorf, MA2, Cristina Villagra, MPH2, Zeida Quintero-Canetti2, Joseph Kannry, MD1, Bruce Darrow, MD, PhD1
1 Clinical Informatics Group, IT Dept., Mount Sinai Health System, 2 Dept. of Pop. Health Science and Policy, Icahn School of Medicine at Mount Sinai, NY, NY

Introduction
The New York State Health Foundation offered OpenNotes (ON) implementation program grants in spring 2017. The Clinical Informatics Group, Mount Sinai Health System (MSHS), was one of seven grant awardees for our proposal to offer a one-year navigator program surrounding our enterprise ON rollout, spring 2018. Our program’s focus was a patient-facing bilingual Spanish-English navigator to help our Spanish-language preferred patient population sign-up for and navigate MyChart accounts, find their ON office visit notes, understand the importance of engaging with their health information, and if desired, provide a verbal translation of clinical notes. Spanish-language preferred patients were the focus because they comprise ~30% of patients at the main Mount Sinai Hospital (MSH) and affiliated practices, which borders East Harlem, in which 50% of the population is Hispanic. This group is the largest with the lowest uptake of MyChart although system-generated numbers on this do not exist.

Methods
From the Jandorf research group, we hired two highly experienced minority-population navigators.1 The second navigator educated staff and clinicians about ON and encouraged them to discuss ON and the navigator program with patients. In spring 2018, we created scripts, FAQs, flyers, signs in Spanish and English; a training PPT; and ordered easel signs. We created question protocols for three rounds of IRB-approved program evaluation; each round comprised two focus groups with navigated patients, two with staff, and interviews with clinicians. Patient navigation launched in May 2018, at our two largest primary care and internal medicine practices. Staff/clinician navigation launched that June, with invitations to all MSH-affiliated outpatient practices. We heavily promoted the program including features in MSHS publications “Your Voice Counts,” the Patient Experience newsletter, the English-Spanish “Our Community” newsletter - distribution of 10,000 including patients; presentations to groups such as the East Harlem Community Health Committee and the Harvard ON group with guests from West Coast systems with large Spanish-preferred populations; hospital lobby tabling sessions; and a Black History Month event.

Results

<table>
<thead>
<tr>
<th>Table 1: End of February Patient Navigation Statistics. MCSU=MyChart Sign Up</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC and ON navigations only</td>
<td>111</td>
<td>21.47%</td>
</tr>
<tr>
<td>MCSU + MC and ON navigation and translation</td>
<td>219</td>
<td>42.36%</td>
</tr>
<tr>
<td>MCSU+ MC and ON navigations</td>
<td>141</td>
<td>27.27%</td>
</tr>
<tr>
<td>ON navigation and translation</td>
<td>35</td>
<td>6.77%</td>
</tr>
<tr>
<td>ON translation only</td>
<td>11</td>
<td>2.13%</td>
</tr>
<tr>
<td>Grand Total</td>
<td>517</td>
<td>100%</td>
</tr>
</tbody>
</table>

We navigated 517 patients. We trained 222 staff/clinicians, ~ three times the goal. Results from the evaluation rounds will be presented. A key finding: Patients loved this program, and because of it, even those with truly limited English felt empowered to use MyChart and ON. Staff/clinicians did not spread the word to patients as hoped, but their ON awareness increased, and they strongly supported the dedicated Spanish-language ON navigator.

Discussion
ON founders at Beth Israel Deaconess Medical Center-Harvard have called our program “ground breaking,” yet from the lessons learned, we know that this program could be recreated readily elsewhere. Where “build it and they will come” fails, the human navigator touch reduces hurdles and empowers patients to succeed.

References
Leveraging a Participatory Informatics Platform for Patient-Reported Outcomes in a Clinical Trial.

Cummings AL, MD; Goldman JW, MD; Arevian A, MD PhD.
David Geffen School of Medicine, University of California - Los Angeles, Los Angeles, CA.

Introduction
The integration of web-based applications for patient-reported outcomes (PRO) has been proven to have significant benefit in cancer care and survival (1, 2). The UCLA Lung Correlative Team (LCT) developed a web-based application, L07-PRO, with the participatory informatics platform Chorus (https://chorus.semel.ucla.edu) (3) to capture PRO for a phase 1b/2 clinical trial, “Talazoparib and low-dose temozolomide in treating participants with relapsed or refractory extensive-stage small cell lung cancer (TRIO-US L07)” (NCT03672773). The trial opened in October 2018 at several sites in California, Nevada, Indiana, New Jersey, and Florida and will enroll 28 participants.

Methods
The design of the application (L07-PRO) was informed by unstructured interviews with and testing by primary investigators, clinical trial coordinators, Chorus staff, and patients. Chorus enables individuals without technical backgrounds to visually create mobile apps through a web interface. A clinical fellow (non-programmer) constructed L07-PRO overseen by Chorus staff. L07-PRO was included as an optional measure and exploratory outcome in the TRIO-US L07 protocol, which received regulatory approval through the UCLA data safety monitoring board (DSMB) and Food & Drug Administration (FDA). The protocol specified L07-PRO assessments be completed at minimum on day 1 of each 28-day cycle, as well as day 14 of cycle 1 and 2. If participants did not wish to complete these assessments through L07-PRO, they could complete them on paper. All entries are user/time/date stamped and trigger email messages if certain criteria are met. Training on the application was scheduled with clinical trial coordinators prior to first enrollment; if participants opted in, they received training during their first study visit. Text messages or emails based on participant preference were sent as reminders the day of the first 4 assessments.

Results
Based on stakeholder feedback, L07-PRO consists of a participant, administrator, and data portal. Once participants opt in, they receive a login and 4-digit pin set through the administrator portal, which is used to review all application submissions. The patient portal includes options to start an assessment, which leads to select multi-choice items from the patient-reported outcomes version of common terminology criteria for adverse events (PRO-CTCAE) (4), report an emergency room or hospital visit, contact their trial coordinator, or provide feedback on L07-PRO or the trial. To date, 3 out of 4 participants have elected to use L07-PRO (age range 40-80s). The patient that declined did not own a smartphone or other mobile device with internet capability. All 3 using L07-PRO requested text message reminders. Of 11 protocol-required submissions, only one was done on paper, related to a forgotten pin. The assessments associated with trial response.

Conclusion
Utilizing participatory informatics in developing PRO applications is feasible in clinical trials. L07-PRO serves as a proof-of-concept for further application development. An adapted version will be used in future lung cancer clinical trials.

References
Missing and Discordant Race Data Documented in Electronic Health Record Systems at an Academic Medical Center

Marika M. Cusick, BA1, Evan T. Sholle, MS1, Marcos A. Davila, MS1, Joseph Kabariti, BS1, Thomas R. Campion, Jr. PhD1,2,3

1Information Technologies & Services Department, Weill Cornell Medicine, New York, NY; 1Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, NY; 3Clinical and Translational Science Center, Weill Cornell Medicine, New York, NY

Introduction

Racial minorities are underrepresented in clinical research, which hinders the ability of investigators and practitioners to address health disparities in vulnerable populations. To identify patients from minority racial groups for participation in studies, researchers can query structured race data documented in electronic health record (EHR) systems. Although meaningful use requirements in the United States require documentation of race in EHR systems for all patients, some patients may have missing data or their race values are discordant across electronic systems. Understanding the degree to which missing or discordant structured race values exist in electronic patient data can inform efforts addressing underrepresented minority participation in research. The objective of this study was to measure patients’ missing and discordant structured race data as documented in multiple EHR systems at one academic medical center.

Methods

We obtained demographics for all patients from the outpatient EHR, inpatient EHR, and inpatient billing system used at Weill Cornell Medicine in New York City. The three systems captured race data according to Federal Office of Management and Budget categories: White, Black or African American, American Indian or Alaska Nation, Asian, Native Hawaiian or Other Pacific Islander, and Other Combinations not described. Because the outpatient EHR contained the most patients, we used it to establish the baseline number of patients, including how many had missing race as defined by NULL or “Declined” values recorded. For patients missing race data in the outpatient EHR, we determined the extent to which race data existed in the inpatient EHR and billing systems. For each patient, we also determined the degree to which race values were concordant across the three systems.

Results

Of 2,887,195 total patients, 1,598,919 patients (53.72%) had missing structured race values in the outpatient EHR. Of these patients, 101,494 (6.34%) had a race value in the inpatient EHR system. Of the 864,742 patients with a structured race value in all three systems, the number of patients with a concordant value across all systems was 547,317 (63.29%) and across two systems 399,743 (35.17%). 22,337 (1.54%) had a differing race value across three systems.

Of the 1,158,702 patients with a structured race value in both the outpatient and inpatient EHR systems, patients identified as Asian (89.62%), African American (88.83%), White (85.72%), and Native Hawaiian or other Pacific Islander (80.06%) patients had concordant races values. In contrast, patients identified as Other Combinations not Described and American Indian or Alaska Nation had higher discordance, as 68.66% and 59.9% had concordant race values, respectively. A Chi-squared test for independence between race and concordant race values proved statistically significant with a value of 85360.62 and an associated p-value of 0.00.

Discussion

We observed missing and discordant data in patient race values from multiple electronic systems within a single academic medical center. Missing and discordant data existed across all races. When researchers and clinicians make use of patient race data, it is important that they are aware of “missingness” and discordance. Informaticians may seek to describe concordance of patient race across electronic systems as an indicator of data quality for investigators seeking to use the values for clinical trial screening and other research purposes. In further analysis, we plan to expand the analysis of concordance with race data obtained from laboratory results and unstructured physician notes using natural language processing.

Acknowledgements/References:

This study received support from NewYork-Presbyterian Hospital and Weill Cornell Medicine, including the Clinical and Translational Science Center (CTSC) (UL1 TR000457) and Joint Clinical Trials Office (JCTO).
Patient Perceptions of Missing Health Information in Outpatient Settings

Elizabeth C. Danielson, MA1, Christopher A. Harle, PhD1,2
1Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana, USA; 2Regenstrief Institute, Indianapolis, IN, USA

Introduction Missing clinical information at the point of care can burden patients, frustrate clinicians, increase costs through repeated testing, and pose safety risks. Although most patients are willing to share their health information across providers, missing information remains a common problem. An estimated 14-49% of patient visits have missing information1,2, while 20-55% have repeat testing3,4. To date, most research and policy has focused on addressing missing information through provider-directed approach. Meanwhile, new policies, like the Centers for Medicare and Medicaid Services’ Merit-based Incentive Payment System (MIPS), incentivize providers to better exchange health information and engage patients. Despite these efforts to engage patients in their personal health record, few studies have evaluated patient perceptions of missing health information or examined whether certain patient populations are more likely to perceive experiences of missing information. To address this gap, we examined two waves of a nationally representative survey to examine the frequency of and factors related to patient experiences of missing information and repeat testing during ambulatory health care encounters.

Methods We conducted a pooled cross-sectional analysis of the 2017 and 2018 Health Information National Trends Survey 5 (HINTS) cycle 1 and 2. Our outcomes were yes/no responses to three questions about patients’ experiences during health care visits: “Had to provide your medical history again because your chart could not be found?”, “Had to redo a test or procedure because the earlier test results were not available?”, and “Had to bring a test an x-ray, MRI, or other type of test with you to the appointment?” We conducted three multivariable logistic regression models to estimate the relationship between each outcome and respondent characteristics: health status, have a usual source of care, survey year, insurance, gender, age, race, disability, household income, education, body mass index (BMI), past cancer diagnosis, smoking status, and Patient Health Questionnaire (PHQ-4) anxiety and depression scores. In order to obtain nationally representative estimates, we applied HINTS survey weights in the models.

Results Approximately 22% of respondents brought a test to an appointment, 6% repeated their medical history, and 6% repeated a test. Patients were more likely to perceive bringing a test if they scored as severe on the PHQ-4 (OR=2.5; p=.002) relative to normal and were in ‘fair’ health (OR=1.9; p=.0017) relative to ‘excellent’. Patients without health insurance (OR=0.45; p=.025) were less likely to perceive bringing a test than patients with private/employee insurance. Patients were more likely to perceive repeating their medical history if they scored as severe on the PHQ-4 (OR=2.0; p=.044) relative to normal and were in ‘fair’ (OR=3.6; P=.023) or ‘poor’ health (OR=6.0; p=.012) relative to ‘excellent’. Patients were more likely to perceive repeating a test with ‘very good’ (OR=2.5; p=.021), ‘fair’ (OR=3.8; p=.003), or ‘poor’ (OR=5.3; p=.008) health status relative to ‘excellent’. Meanwhile, patients with incomes in categories greater than or equal to $20,000 were less likely to perceive repeating a test than patients with incomes less than $20,000 ($20k-49k OR=0.41, p=.005; $50k-74k OR=0.44, p=.048; >$75k OR=0.43, p=.039).

Conclusion Patients perceive missing health information and repeat testing much less often than estimates from studies of health records. Moreover, patients in some vulnerable groups, such as those without health insurance, low income, or poorer health status may be more likely to perceive they have experienced missing health information or repeated tests. These findings suggest an opportunity to increase patient awareness of the frequency of missing health information, including targeting awareness toward patient groups that are more likely to experience missing information.

References
Acceleration Gait Measures as Proxies for Motor Skill of Walking: A Narrative Review
Pritika Dasgupta, MPH MHI MS; Jessie VanSwearingen, PhD; Alan Godfrey, PhD; Ervin Sejdić, PhD
University of Pittsburgh, Pittsburgh, PA

Introduction

Acceleration gait measures (AGMs) are a promising component to motor skill research, which can help older adults’ (65 years or older) quality of life and reduce the burden of fall-related healthcare. A decline in motor skill or control\(^1\), often associated with both cognitive and gait changes during aging, is a public health matter because it is one of the significant causes of falls, morbidity, and low quality of life\(^2\). However, motor skill is multifaceted and complex. For example, older adults may walk slowly with adapted optimal motor skill some may walk slowly with poor motor skill, and some older adults with or without diagnosed disease may walk at clinically normal speeds with altered control. The first line of gait assessments is known as collecting simple gait measures, which researchers collect through observation, magnetic systems, optical systems, and video recording; these measurements consist of visual observations of spatiotemporal gait measurements, which include gait speed as well as stride and step characteristics\(^3\). Unfortunately, research suggests that simple gait measurements are not able to contradistinguish between age and disease-related gait dysfunction\(^4\), and simple gait measurements do not provide adequate information about the quality of the translation of the body motion during walking. AGMs can enrich simple gait measurements of walking and motor performance by allowing researchers to see if the motor skills were compensated (such as energy expenditure or smoothness).

Methodology

Research literature was drawn from published journal articles and conference abstracts in the following scholarly databases: PubMed, IEEE Xplore, and SpringerLink. This narrative review does a critical analysis of categorizing motor skill aspects of walking and organizing AGMs. To date, there has not been a wide-scale mapping of specific motor skills aspects of walking to AGMs – we seek to survey the literature to understand how AGMs are being used as proxies for motor skills.

Motor Skill

The motor skill of walking results in smooth and efficient translation of the body over the surface. A decline in motor skills refers to a loss of the coordination and timing of stepping with postures and phases of gait in the performance of walking. Based on the literature search, we defined seven characteristics of the performance outcome of motor skill of walking: 1) smoothness, 2) efficiency, 3) automaticity, 4) adaptability, 5) variability/regularity, 6) stability, and 7) symmetry.

AGMs

AGMs are one of the multiple ways to measure human movement, but other measures are costlier, demand challenging environment setups, and are bulky. Based on how researchers perform signal pre-processing and the nature of their research questions, there are many different forms of AGMs\(^4\). We organized these AGMs via various signal pre-processing methodologies/groupings: 1) gait cycle timings, 2) statistical features, 3) signal-frequency features, 4) time-frequency features, and 5) information-theoretic features.

Mapping Motor Skill and AGMs

Clinically, mapping motor skill characteristics to categories of AGMs may be capable of providing relevant and accurate measurements. In the review, we summarized a selection of references for each of the motor skill-AGM mappings. By doing so, we will address the existing gap by seeing how researchers have combined multiple features extracted from gait accelerometry signals into a derived AGM that will potentially be a marker for walking related changes in physical function.

Conclusion

We have an urgent need for more holistic motor skill research since people are living longer and may be more prone to neuromotor illness; for example, fall-related injuries result in millions of US patients per year in emergency rooms; for example, over 800,000 patients/year are hospitalized\(^5\). In application, AGMs have been used and can be used more effectively to detect differences and changes in motor performance due to learning/expertise, or task and environment manipulations\(^6\).

References

Teledermatology service in HealthNet Platform: an Experience Report

Keilla T. Martins de Melo, M.Sc.¹, Nelson G. de Sá Leitão Júnior, M.Sc.¹, Danilo F. S. da Silva, M.Sc.¹, Webber S. Fantini, M.Sc.¹, Alexandra M. M. Grisolia, Ph.D.² Magdala de A. Novaes, Ph.D.¹

¹Telehealth Center (NUTES), Federal University of Pernambuco, Recife, PE, Brazil.
²Biomedical Center, State University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil.

Introduction

A teledermatology service is a dermatological practice performed without the need for face-to-face intervention of a specialized healthcare practitioner. This practice, which is a form of telediagnostic, has a relevant visual component that favors the development of iconographic representation techniques in telehealth. Teledermatology aggregates the documentation of clinical information and images of dermatological lesions for healthcare, offering additional benefits to the traditional methods of dermatological evaluation. The Telehealth Center (NUTES) of the Clinics Hospital of the Federal University of Pernambuco (UFPE) implemented a set of specific features on the HealthNet software platform to support the Teledermatology service, which is the object of this case report.

Specific purposes of the system

The HealthNet¹ is a telehealth platform that offers integrated teleconsulting and telediagnostic services, facilitating the clinical cooperation and the exchange of experiences among healthcare practitioners from the Brazilian government Unified Health System (SUS). In this context, the NUTES team implemented a set of specific features for practice, based on the requirements of contextualization of clinical practice, the incorporation of medical images and the easy access to information as a tool to support a medical decision. Those requirements were constructed according to archetypes that were applied in the subdivisions Action, Instruction, Observation, and Evaluation; grouped into archetypes of the Section type². Each subdivision corresponding to an activity related to teledermatology, such as Action for the execution, Instruction for the request, Observation for the description of aspects observed during the execution, and Evaluation for the exam results.

Results of the usage

In telediagnostic, a health care practitioner, as a solicitant, submit a clinical exam in the HealthNet for obtaining this exam results from a remote health care specialist. In our teledermatology service, a solicitant submits a particular telediagnostic with the patient information, additional exam data, and specific dermatological data by using a specific HealthNet feature for the dermatoscopy procedure. The solicitant includes a structured medical record, according to the mentioned archetypes, and he may also include detailed lesion information with the help of a graphical human body model delimited by dermatologic regions and images of the patient, both for further analysis. When the solicitant concludes the exam submission, another healthcare practitioner, identified as a teleguider, redirects this telediagnostic request to an available health care specialist, who in turn, may request more details on this matter to the solicitant or directly perform an analysis on the exam data to submit the results in the platform.

Conclusion

Based on our experience with the HealthNet platform and the results from remote specialists concerning dermatoscopy exams, we conclude that adopting teledermatology is a viable approach for healthcare practitioners. Additionally, we may highlight that in the last six months of pilot operation, practitioners performed more than 140 dermatoscopy exams by the HealthNet, thereby, fully supporting the proposal of the teledermatology service.

References


Federal Research Data Infrastructure: Tools and Services for Health Care Needs

Rina Dhopeshwarkar, a MPH, Lauren Hovey, a MA, Krysta Heaney-Huls, a MPH, Prashila Dullabh, a MD, Scott Smith, b PhD

a NORC at the University of Chicago, Bethesda, MD; b The Office of the Assistant Secretary for Planning and Evaluation, Washington, DC

Research Objective

The proliferation of electronic health data is a rich resource for the health system, with the potential to fuel research and evidence generation that expands knowledge about the effectiveness of health interventions. Challenges have emerged, given the volume and technical complexity associated with data collection, aggregation, and analysis. A variety of tools, standards, and services are needed to make diverse electronic health data available and usable for patient-centered outcomes research (PCOR). NORC at the University of Chicago analyzed a sample of projects funded by the PCOR Trust Fund (PCORTF) to identify solutions and the types of challenges they ameliorate.

Study Design

The Patient Protection and Affordable Care Act (ACA) created the PCORTF to build national data capacity and infrastructure and to help researchers leverage existing data for PCOR. It is administered, in part, by ASPE. We conducted a retrospective analysis of 11 projects funded by the Office of the Assistant Secretary for Planning and Evaluation (ASPE) whose activities concluded in 2018. We then categorized the projects’ contributions to broader health system issues and research goals. Our review included project statements of work, quarterly reports, meeting summaries with project leads, final reports and deliverables. This retrospective approach allowed us to assess project goals and trajectories, outputs (e.g., services and standards), and their broader significance to the health system.

Principal Findings

The 11 projects, conducted at six federal agencies, addressed PCOR data infrastructure needs and innovations in the following categories: 1) use of application programming interfaces (APIs) to improve capture and use of health data; 2) enhancements to data governance structures; 3) patient matching solutions; 4) facilitators for research across distributed networks; 5) data standardization; and 6) improved mechanisms for linking patient records. These projects also addressed federal strategic priorities, including: 1) collection of participant-provided information; 2) standardized collection of standardized clinical data; 3) linking of clinical and other data for research; 4) use of clinical data for research; and 5) use of enhanced publicly-funded data systems for research.

Conclusions

The PCORTF projects contribute substantially to PCOR data infrastructure, addressing multiple priority area. They do so via an array of services, standards, analytic and patient matching tools, and governance structures that facilitate the exchange and use of electronic data from diverse sources. As such, their products are of practical use to PCOR stakeholders and many are publicly available for immediate use.

Implications for Policy or Practice

The health system goals to support patient and provider decision-making with evidence-based medicine, to improve patient outcomes, and address pressing health issues like opioid use require robust health data infrastructure. The PCORTF projects offer a variety of research-focused tools for aggregating, linking, and analyzing data; as well as tools designed to increase patient autonomy in managing their health records and contributing data to research, all of which have broad potential for use and utility in PCOR.


3 Assistant Secretary for Planning and Evaluation (ASPE). The OS-PCORTF Strategic Framework. https://aspe.hhs.gov/os-pcorn-strategic-framework
Patient Perspectives of Utilizing Multiple Mobile Health Technologies to Self-manage Type II Diabetes

Anna Diane BSN RN1; Allison Lewinski PhD2; Jacqueline Vaughn BSN RN1; Janee Stevenson BSN, RN3; Ryan J. Shaw PhD RN1

1Duke University School of Nursing, Durham, NC; 2 Durham Center of Innovation to Accelerate Discovery and Practice Transformation, Durham VA Health Care System, Durham, NC 3 Winston-Salem State University, Winston-Salem, NC

Introduction: Type II Diabetes (T2D) is a complex, chronic illness that requires daily self-management of blood glucose levels, weight, medication adherence and attention to lifestyle behaviors. Mobile health (mHealth) technologies (e.g., wearables, wireless glucometer, wireless scale, text message surveys) facilitate the collection of real-time, in situ patient-generated health data (e.g., physical activity, blood glucose, weight, and medication adherence)1. These data can help patients monitor their daily behaviors thus improving the patient’s ability to engage in diabetes self-management. These real time data can be leveraged to deliver real-time targeted self-management interventions.

Methods: We conducted a mixed-methods exploratory study to examine the feasibility and utility of patients using multiple mHealth devices in diabetes self-management. We provided participants (N=60) with a wearable accelerometer, wireless glucometer and wireless scale during the 6-month study period to respectively track physical activity, blood glucose and weight. We purposively selected 20 patients based on study engagement (e.g., high, low, medium) to participate in semi-structured interviews. The interviews queried patients to discuss their experiences in self-managing diabetes with the three mobile devices.

Results: Directed content analysis was performed and analyzed to describe participants’ perceptions on utilizing mobile health technologies to collect and monitor their diabetes related health data. We identified two main themes related to the patients’ perceptions of utilizing mHealth devices in diabetes self-management. The first theme, feasibility, described how easy or difficult the patients found each device to use. We noted two sub-themes: (1) usability—which described the extent to which the devices were used to manage diabetes-related tracking behaviors; and (2) satisfaction—which discussed the level of acceptability of the devices. The second theme, utility, described how they used each of the three devices over the six-month study period. We noted four sub-themes (1) accountability—which discussed prompts and reminders for self-management behaviors; (2) facilitation of conversation—which detailed conversations that participants had with others about device and data utility; (3) accuracy—which described the patient’s perceptions regarding the validity of devices and; (4) accessibility—which discussed the participant’s ability to have their own data within reach.

Conclusions: Our findings indicate the importance of obtaining patient feedback in utilizing multiple mHealth devices to self-manage diabetes. Understanding the range of participant-centered perspectives can inform design of future mobile health interventions aimed at promoting patient self-management of chronic illnesses.

References:
Ascertaining Medication Adherence Utilizing Open Claims Data

John E. Doole, Pharm.D., MFA1, Pamela B. Landsman-Blumberg, MPH, DrPH1, Matvey B. Palchuk, MD, MS, FAMIA1,2

1TriNetX, Inc., Cambridge, MA; 2Harvard Medical School, Boston, MA

Introduction

Comparative effectiveness studies, predictors in medical outcomes models, and targeting patients to medication therapy management programs are typical applications of medication adherence metrics. Adherence is usually calculated using “closed” claims data. These are fully adjudicated and contain data for all prescription fills available within a fixed period. A different source of claims data – “open” claims – is used much less frequently to measure adherence. These claims cover a much broader swath of payers, may not be fully adjudicated and may be missing claims for fills from pharmacies utilizing a clearinghouse not included in the pool of data providers. We set out to show that medication adherence can be determined using open pharmacy claims data.

Methods

Using a commercially available open pharmacy claims dataset covering 99% of US payers, we calculated common measures of medication adherence. Data elements such as supply duration, quantity dispensed, fill number and corresponding date ranges are needed to determine Medication Possession Ratio (MPR), which measures the percentage of time a patient has access to a particular medication. Given the potential for an incomplete view of the claims lifecycle, we applied a number of heuristic rules to de-duplicate records for a given prescription fill where correction and denied claims may be submitted before the clearinghouse sends a final claim to the payer and a final remittance is sent back to the clearinghouse and ultimately the dispensing pharmacy.

Results

We calculated MPR for all drugs in the dataset. Adherence with MPR of 0.9 and above was considered high, 0.8-0.89 – fair, and under 0.8 – poor. We evaluated our calculations by visualizing the relationship between prescription complexity (>2 takes per day=complex) and route of administration to adherence. The results are in line with our expectations. A Sankey diagram (see Figure 1) demonstrates, for example, that lower prescription complexity leads to better adherence and that patients are more compliant with oral medications.

Conclusion

Open claims datasets pose unique challenges for calculating medication adherence. We demonstrated that it is possible to generate such metrics and confirmed our results by seeing expected trends in relationships between prescription complexity and the route of administration to medication adherence.

References

Improving Accuracy of Patient Speech Transcription Using Dialect-Specific Automatic Speech Recognition Models

Rachel Dorn, B.A.¹, Scott Vrana, Ph.D. ², Bridget T. McInnes, Ph.D.¹
¹Computer Science Department, Virginia Commonwealth University, Richmond, VA, USA
²Psychology Department, Virginia Commonwealth University, Richmond, VA, USA

Introduction. Much progress has been made in automatic speech recognition (ASR), but it has not been evenly distributed across all dialects of English¹. African American Vernacular English (AAVE) is a widely-spoken English dialect in the United States, yet it is grossly underrepresented in common speech corpora². As a result, patients who speak this dialect are at risk of being misunderstood by clinical speech applications more often than Standard American English (SAE) speakers, negatively affecting their user experiences and hindering downstream clinical applications that rely on ASR data. This study explores the effect on transcription accuracy of using AAVE data.

Methodology. The dataset consists of 36 AAVE-speaking interviewees and 6 SAE-speaking interviewers from the Corpus of Regional African American Language (CORAAL)³, which were split into 31,471 utterances by speaker. Transcription models for AAVE-only utterances, SAE-only utterances, and the combined set of the two (Combined) were trained using Mozilla’s DeepSpeech ASR engine⁴. We evaluated the models using word error rate (WER) and Levenshtein distance. Levenshtein distance was included as a way to reward partial credit, as many of the outputs included correct syllables but incomplete words. Lower numbers indicate higher accuracy for both measures.

Results. Table 1 shows that both the average WER and Levenshtein distance decreased for both AAVE and SAE models as compared to the Combined model. The WER rates are nearly all above 1 due to the outputs putting in extra word breaks. There was an 18% improvement across both metrics for the AAVE utterances and a 13% improvement for SAE utterances. The larger increase for AAVE is likely due to the combination of: 1) AAVE contains a larger number of utterances in the dataset; and 2) the average length of an AAVE utterance is longer than an SAE utterance.

<table>
<thead>
<tr>
<th>Data</th>
<th># Utterances</th>
<th>Average Utterance Length</th>
<th>Model</th>
<th>Word Error Rate</th>
<th>Levenshtein Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAVE</td>
<td>22,587</td>
<td>29.372 characters</td>
<td>AAVE</td>
<td>0.93628</td>
<td>15.28477</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.15086</td>
<td>18.75011</td>
</tr>
<tr>
<td>SAE</td>
<td>8,884</td>
<td>21.749 characters</td>
<td>SAE</td>
<td>1.04134</td>
<td>12.99667</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1.12545</td>
<td>15.05882</td>
</tr>
</tbody>
</table>

Conclusions. Different dialects within the same language can have conflicting patterns of phonemes; the ambiguities arising from having to learn opposing patterns can cause errors and lower overall accuracy. Applications which handle diverse patient speech can benefit from using dialect-specific models for speech recognition. This is particularly useful in contexts where the speech data is gathered in advance, rather than in real-time. Future research will train a classifier to automatically detect the dialect being spoken as well as to expand the datasets to improve overall accuracy.

References
Customization of an Antimicrobial Stewardship Clinical Decision Support Module: Reducing the Noise and Improving Reporting

Evan Draper, PharmD¹, Brian K. Fung, PharmD¹, James Steckelberg, MD¹, Abinash Virk, MD¹, Sara Ausman, PharmD², Sara Lessard, PharmD³, Aaron Tande, MD¹, Lynn L. Estes, PharmD¹

¹Mayo Clinic, Rochester, MN, ²Mayo Clinic Health System, Eau Claire, WI, ³Mayo Clinic Health System, La Crosse, WI

Introduction
Mayo Clinic recently implemented Epic® as a single Electronic Health Record across all of Mayo Clinic’s campuses. Implementation of Epic® replaced two Antimicrobial Stewardship Clinical Decision Support (CDS) systems. Antimicrobial Stewardship CDS systems use rules to flag patients for clinical review. Logic within a CDS system cannot replace clinical judgement and some alerts will be non-actionable.¹ One of the replaced systems was a mature, home-grown system² that allowed users to defer or dismiss rules after review and provided reports on the frequency of alert firing. Current limitations in the out-of-the-box Epic® Antimicrobial Stewardship module include an inability to turn off rules after review and limited reporting on alert firing.

Methods
We developed a novel approach to surveillance within Epic® that allows users to defer or dismiss flags and provides the ability to report on when rules fire. During a pilot of the novel approach from 1/1/18 to 2/28/18 we captured the number of days any patient is flagged for review at 7 pm for the usual approach and the novel approach. The rules contained the same clinical criteria but the novel approach allowed users to dismiss or defer the rules.

Results
Over 2 months a comparison of the approaches shows a reduced frequency of alerting with the novel approach.

<table>
<thead>
<tr>
<th>Rule Description</th>
<th>Usual Approach (#{days alert is active})</th>
<th>Novel Dismiss/Defer approach (#{days alert is active})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antimicrobial timeouts</td>
<td>496</td>
<td>241</td>
</tr>
<tr>
<td>Positive culture and no antimicrobial</td>
<td>52</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 1 compares the number of days an alert is active between the usual and novel approaches:

Discussion: Customization of the Epic® Antimicrobial Stewardship module resulted in fewer recurring alerts and improved reporting.

References
Clinical Decision Support Flow Diagram for Management of Chronic Kidney Disease by Primary Care Physicians

Vishal Duggal, MD1,2, Mary K. Goldstein, MD, MS1,2

1 VA Palo Alto Health Care System, Palo Alto, CA
2 Stanford University School of Medicine, Stanford, CA

Chronic kidney disease (CKD) affects more than 15% of adults in the United States1, and the Veterans Affairs health system is the single largest provider of health services in the U.S. for patients with CKD2. The MEDSAFE-CDS is a knowledge-based system developed jointly by the Department of Veterans Affairs (VA) and Stanford University that is designed to aid primary care physicians within the VA Health Care System in the identification and management of multiple chronic health conditions. This clinical decision support (CDS) tool, currently in development for MEDSAFE, provides direction for physicians caring for patients with CKD Stage 3 and guideline-supported recommendations to manage the progression and complications of CKD.

CDS systems require specific disease definition criteria to identify eligible patients. While phenotypic determinations exist for CKD, they do not identify groups specifically amenable to care via CDS. Here we present a preliminary decision flow diagram that adheres to VA/Department of Defense (DoD) Clinical Practice Guidelines for the Management of CKD in Primary Care. This diagram details the entry point into the CDS tool, decision branch points, disqualifying criteria, and recommendations for physicians including laboratory tests, medication prescriptions, and conditions that should prompt referral to specialist care. Recommendations from this CDS are provided in an online report for primary care physicians alongside recommendations for chronic diseases such as hypertension and diabetes mellitus. Such reports can be used in pre-clinic huddles and during patient visits.

The CDS decision flow diagram, a portion of which is illustrated in Figure 1, specifically includes recommendations for frequency of kidney function laboratory testing, assessment of estimated glomerular filtration rate trends, management of metabolic acidosis, and investigation of anemia related to CKD. The VA/DoD guidelines were augmented with those of Kidney Disease International Improving Global Outcomes (KDIGO)3.

Figure 1. Subset of the ATHENA-CDS Clinical Decision Support Flow Diagram

Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs.

Leveraging Technology to Promote Healthy Cooking Literacy and Kitchen Confidence for Overweight and Obese Adolescents

Polina Durneva1, Cynthia LeRouge2, Savitha Sangameswaran3

1Florida International University, Miami, FL; 2Florida International University, Miami, FL; 3University of Washington, Seattle, WA

Introduction

Obesity and overweight are a major public health issue affecting 12.7 million children and adolescents. One of the recommendations to combat obesity is to maintain healthy diet and eat home cooked meals. Providing user-centered technology-enabled solutions to overweight adolescents can equip them with necessary skills towards managing their health behaviors. Most of commercially available apps, however, do not address all components of Culinary Nutrition Education (CNE) (e.g. knowledge about intake of fruits and vegetables, sodium, and added sugar) and are not tailored to the adolescent audience. In this design science study, we therefore aim to inform functional requirements, design requirements, and insights regarding contextual use of tailored CNE technology that could enhance teens’ healthy cooking literacy and kitchen confidence - thereby promoting their healthy lifestyle management through personal choices, increased self-efficacy, and influencing family eating habits.

Methods

The study was completed in 2 distinct phases that engaged overweight and obese adolescents, parents, and providers. Phase 2, our primary source of data, consisted of multiple user-centered design methods (e.g., usability analysis and semi-structured interviews) involving mid-range prototypes informed by Phase 1 with 70 adolescents from a healthy behavior camp and 10 providers. Data were analyzed using the constant comparative method analysis to identify functionalities, design preferences, and context of use of our technology.

Results

Figure 1 summarizes findings regarding functional and design requirements as well as the context of use (environment, motivators, and barriers) of a CNE technology.

Discussion

The study reveals design and functional requirements of the CNE technology that reflect constructs from Social Cognitive Theory, one of the main behavior change theories focusing on the ongoing change in person’s behavior, and the Fog Behavior Model (FBM), which explores the effect of persuasive design on people’s behaviors. Our findings demonstrate support and provide contextual insight for these theoretical models.

Conclusion

Our technology-proposed solution reflects the goals and identified requirements of overweight adolescents; it also aligns existing behavior change theories to this technological context. We demonstrate the potentiality of creating a technology that could effectively lead to better self-care, induce sustainable behavioral change by way of education and activity, and promote self-efficacy by enhancing cooking literacy and kitchen self-confidence.
Understanding Nothing: Interpreting Models Using Sparse Longitudinal Data

Jacob M. Ekstrum B.S.¹, Noah J. Pocziwinski B.S.¹, Brett F. Cropp M.S.¹, John B. Coles Ph.D.¹
¹CUBRC Inc. Information Fusion Group, Buffalo, NY, USA

Abstract
Recent improvements in machine learning have led to the ability to generate both complex and sophisticated time-series models for healthcare situations. Though powerful, these models traditionally lack insight into the importance of each of the model’s input features. Here, we introduce a novel element of the Heartwood Analytics™ system that ranks both sparse and dense time-series features according to their importance in the model for clear communication of driving factors for complex medical decision making.

Introduction
Models integrated into important aspects of an individual’s healthcare must be explainable, lest the growing trust between professionals and computer systems begin to wane. Current models that can explain their decision process, such as linear models and shallow decision trees, lack the accuracy that attention-oriented methods, such as LSTM-based recurrent neural networks, produce for sparse longitudinal data. Even for those models that can be interpreted, it is difficult to concisely convey important information to care providers and clinicians as the input variables to be ranked increase in both count and sparsity, as is often the case with complex healthcare scenarios. Other attempts to promote interpretable models, such as the boosted tree mimic model described by Che et al.¹, use methods that are interpretable at their higher layers but become much less interpretable as the mimic model is explored.

Determining Insights
The Heartwood Analytics™ system addresses these problems through its feature importance framework. The data used to train the model is acquired and pre-processed, pruning sparse information and concatenating sparsity meta-information for each instance that heavily impacted model training. With sparse data condensed to a dense format, feature importance uses non-linear transforms to acquire an alternative representation of the selected input data. In a new data space with synthetically generated model features, this modified dataset is used to train an interpretable model, such as a linear model. This provides insight into how important each synthetic feature is to the complex model being evaluated. By reverse mapping the importance values for the synthetic features into the original feature space through an approximate inverse function, the framework can generate an accurate and easily interpretable scalar ranking value for each input variable, input timestep, and sparsity meta-feature in the complex model. This mapping procedure is designed to improve accuracy over traditional meta-model approaches while maintaining the interpretability that linear meta models provide. The global approximations embedded in our meta-model are used to generate per-patient models of the impact of variables, which are gaining popularity among Heartwood adopters in the complex care and autism care spaces.

Explaining Variable Impact
Once these scalar values are obtained, Heartwood Analytics™ performs one additional step by explaining and communicating the results to less-technical audiences, such as clinicians and staff that require more information to reduce autism-related behavioral events. Through a custom web interface, features along with their scalar importance values and directionality (positive or negative impact on outcome) are displayed to the user. Adjustments made to feature values in the interface generate real-time predictions alongside previous predictions in order to communicate feature importance. Allowing healthcare providers to verify a model’s accuracy through medical intuition builds trust in the model’s ability to improve decision making overall, supporting better patient outcomes.

Conclusion
By leveraging high-dimensional global interpretability measures, the Heartwood Analytics™ feature importance framework provides an interpretable ranking scale that input variables can be measured against at different points in time. Through the use of a custom web framework, these generated rankings can be filtered in order to tactfully communicate high-impact insights from the model to care providers in myriad health-related situations.

References
Challenges to a Data Driven Approach to Population Level Analysis of Hypersensitivity Events in Cancer Clinical Trials

Christina Eldredge, MD¹, James E. Andrews, PhD¹. Maryam Zolnoori, PhD², Timothy Patrick PhD³, Joel Gallagher, MD⁴, Cesar Lam, MD⁵, and Jake Luo, PhD³

¹University of South Florida, Tampa, FL; ²Mayo Clinic, Rochester, MN; ³University of Wisconsin-Milwaukee, WI; ⁴Allergy and Asthma Center of North Carolina, Greensboro, NC; ⁵Moffitt Cancer Center, Tampa, FL

Objective: To present an analysis of the terms used to report hypersensitivity events (HEs) in cancer clinical trials in order to better understand how HEs vary across studies.

Introduction: National efforts to accelerate cancer research and improve outcomes have led to improved patient access to innovative therapies¹. Yet, new therapies also increase risks for adverse events, including HEs. HEs may present as a non-serious event (e.g. mild allergic reactions such as rhinorrhea) or a serious event (e.g. life-threatening severe allergic reactions, referred to as anaphylaxis). Coding of these events using standard terminologies is complicated and inconsistent, especially given the lack of universal agreement on a specific definition of “anaphylaxis”²; however, effective use of terminologies for representing HEs is needed for data sharing, reuse, and interoperability across cancer trials.

Methods: Investigators are required to submit clinical trial adverse events to ClinicalTrials.gov in the form of two tables, serious and other-than-serious events. ClinicalTrials.gov specifies how data elements related to adverse events (including HEs) are to be entered³. Investigators select a Collection Approach (either Systematic Assessment or Non-Systematic Assessment) and a standard terminology⁴. A publicly available dataset of cancer clinical research trial outcomes and adverse events reported to ClinicalTrials.gov for the period November 1999 to November 2017 was analyzed. HEs were extracted from semi-structured text, within adverse event tables, using an XML parser. We then determined whether investigators used systematic or non-systematic data collection methods, and which standard terminology they used to characterize HEs.

Preliminary Results: Preliminary analysis of 582/5,596 HEs from 1,331 clinical trials revealed the majority of events, 366/582 (63%), were not recorded using a systematic method, and in 64/582 (11%) of HEs no method was specified. Variations were found across clinical trials with several types of terminologies used to code HEs including: COH, CTC (v3 and v4), CTCAE (2.0, 3.0, and 4.0), and MedDRA (versions 5, 9, 10, 12, 13, 15, 17-20). Term inconsistencies were evident when coding severe allergic-type hypersensitivity reactions, especially with regard to coding the etiologic agent. For example, the following terms and their respective vocabularies were used: Anaphylaxis (CTCAE and MEDRA); Anaphylactic shock (MEDRA); and, Bactrim anaphylaxis (Not specified). Anaphylaxis can occur with or without shock symptomatology, so it is unclear whether these events are indeed equivalent. Additionally, term use inconsistencies were also seen within clinical trials; for example, as the result of using several versions MedDRA to code “anaphylaxis” within a single clinical trial.

Future Work: Our analyses suggest barriers exist to population level analysis of HEs in cancer trials, including an overall lack of granularity when coding the severity of these events, as well as a lack of consistent notation of the etiologic agent. We plan to normalize the identified HEs by mapping them to SNOMED CT, which has better coverage for allergy reporting. Normalized data can be used to evaluate the association between cancer intervention therapies and HEs.

References

Towards Optimizing Usability in e-Government Healthcare Administration Systems: Recommendations from a Comprehensive Evaluation Study

Forhan Bin Emdad, BSc, and Güneş Koru, PhD
University of Maryland, Baltimore County, Baltimore, MD

Introduction
Several research studies have reported that perceived system complexity and a lack of usability prevent e-government healthcare administration systems from achieving their full potentials.1 In spite of the presence of general usability guidelines and considerable investment of resources for design and development, usability problems are still persistent in e-government healthcare administration systems.2 It is still challenging to guide government healthcare and information technology(IT) professionals in a right direction in this domain. Therefore, this study had the following objectives: (a) applying a usability evaluation initiative to understand and document top usability challenges and improvement opportunities in a real-life e-government system, and (b) provide a set of recommendations that can be useful for the government officials and IT vendors which can further benefit similar e-government healthcare administration systems.

Methods
We have designed and conducted an iterative usability evaluation and improvement project for an eminent active e-government healthcare system adopted by one of the states in the US. In the first iteration, seven usability experts performed an evaluation of the system’s usability compliance with 10 heuristics.3 Secondly, person-to-person interviews using think-aloud protocol was conducted with 23 participants for usability testing. Thirdly, an online survey based on the questions in the System Usability Scale (SUS) was conducted with 1,800 system users for a quantitative assessment of usability. For analyzing user satisfaction in SUS, maximum SUS score was set to 100 with three categorizations: above average (80 or higher), average (68 to 79), poor (52 to 67), and unacceptable (51 or lower) based on usability condition. Finally, a semi-structured focus group interview was conducted with eight participants including government officials working in health care administration system to obtain further detailed qualitative information regarding usability challenges and improvement opportunities. Accumulated quantitative data were analyzed by statistical techniques and qualitative data were analyzed by transcribing interview records, and triangulation.

Results
The top usability challenges identified were categorized using weighted metrics considering priority and severity.4 The usability experts identified 41 unique usability challenges through heuristic evaluation of the system. Among them, top challenges were navigation problems, unfamiliar icons, and a complex index page structure. Nine additional usability problems were found from person-to-person think-aloud interviews where confusing name conventions, lack of tutorials, and lack of filtering options were identified as the top usability challenges. The survey resulted in a SUS score of 35, which is a very low score for any system, indicating that usability problems should receive attention. Furthermore, in online surveys, almost 20% participants "strongly disagreed" that learning this system is easy, instructions are understandable, system performs without error or system is comfortable to use whereas nearly 13% strongly agreed. Over 20% participants had neutral point of view. In focus groups, challenges such as inexperienced developers, lack of training for the users, and lack of co-ordination among users, developers, and stakeholders were most highlighted for improving usability of the e-government healthcare administration system. Initiatives can be taken to make proper investments to hire skilled IT professionals to resolve these challenges and to train the system users to perform their task appropriately.

Conclusion
This study has successfully documented top usability challenges and improvement opportunities focusing on a real life usability evaluation project for an e-government healthcare administration system. Furthermore, this study has drawn attention to the potential benefits of making recommendations based on the usability challenges identified in this study and other similar studies in the future. In the future, researchers should study the people and organizational challenges of usability improvement in e-government administration systems.

References
Preliminary chart review for Natural Language Processing development among electronic health records with documentation of marijuana use

Termeh M. Feinberg, PhD, MPH1,2, Joseph Goulet, PhD1,2, Amy Justice, MD, PhD2,3, Sameh Fodeh, PhD1,2, Lori Bastian, MD, MPH2,3, Qing Zeng-Treitler, PhD4, Cynthia Brandt, MD, MPH1,2

1Yale University School of Medicine Center for Medical Informatics, New Haven, CT; 2VA Healthcare Connecticut Pain Research, Informatics, Multimorbidities, and Education (PRIME) Center, West Haven, CT; 3Yale University School of Medicine Department of Medicine, New Haven, CT; 4George Washington University Biomedical Informatics Center, Washington, DC

Abstract
Marijuana use is increasing in the U.S., due to increased legalization for medical use and may be used by patients with pain. The development and refinement of search terms is a necessary step preceding use of informatics tools to identify marijuana documentation in the free text of patient charts. We used low-level text processing tasks to compare terminology used in a different sample. Documentation of any positive-use term was present in 56% of pulled charts.

Introduction
Marijuana is the most common illicit substance used worldwide with increasing use in the U.S. due to increased legalization for recreational (10 states) or medical use (33 states) and may be used by patients with a musculoskeletal disorder as a maladaptive coping strategy. Epidemiological examination of marijuana in relation to health outcomes remains sparse, and there is a lack of rigorous evidence for both benefits and risks. Unstructured text (i.e. ‘patient notes’) within Veterans Health Administration (VA) Electronic Health Records (EHRs) are largely untapped in relation to marijuana use among Veterans in VA care. The development and refinement of search terms is a necessary step preceding use of informatics tools to extract, retrieve, and automate data extraction on patients who report marijuana use. We sought to replicate and compare the lexicon used by colleagues in a separate published research study, while assessing presence of the term ‘medical’.

Methods
We conducted a manual chart review of 75 EHR notes chosen from 1000 randomly selected veterans diagnosed with musculoskeletal disorder who had any documented encounter with a VA provider in any clinical setting. We categorized EHRs into those containing information on marijuana use in the past or present, had either positive or negative medical marijuana classification, and documentation containing lexicon terms with no meaningful information regarding endorsement/use. Our lexicon contained the following: “marijuana”, “mjx”, and “cannabis”.

Results
Overall current documentation of any term was present in 56% of EHRs (42/75); current use of “marijuana” was present in 24 EHRs, in addition to adjacent term “medical” in 7; medical-related “cannabis” was not referenced. A positive current use of “cannabis” was documented in 21 EHRs, with few (3) duplicates between “cannabis” and “marijuana” and 1 EHR with conflicting documentation across both terms. Overall previous use was documented in 19% of EHRs (14/75); the term “marijuana” was used to indicate previous (i.e. past) use in 10 EHRs while previous use of “cannabis” was present in 6 (2 overlap). The term “mjx” was not contained in any EHRs, while 4 contained terms not resulting in meaningful information (i.e. unfilled templates(2), mention of cannabis cultivation(1), and provider instruction not to engage in use(1)). Negation statements regarding a lack of current use was minimally present for both “cannabis” and “marijuana” (5 and 3, respectively).

Conclusion
“Marijuana” was the preferred term when specifically documenting non-VA medical marijuana. There was minimal overlap between “cannabis” and “marijuana”, suggesting both terms may be uniquely important for lexicon inclusion. Providers should continually update EHR templates and outdated notes in relation to patient marijuana endorsement in order to create meaning from data and turn new knowledge into action. Future research should continue refinement of identification of marijuana use within EHRs by inclusion of additional terms.
Aiding in the Identification of Acute Respiratory Distress Syndrome (ARDS) from Radiograph Reports using Natural Language Processing

Jeffrey P. Ferraro PhD1,2, Jason Jacobs, PhD3, Ithan Peltan, MD3, Michael Lanspa, MD3, Raj Srivastava, MD4,5, Colin K. Grissom, MD3

1Intermountain Healthcare, Salt Lake City, Utah; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah; 3Intermountain Medical Center, Division of Pulmonary and Critical Care, Salt Lake City, Utah; 4Healthcare Delivery Institute, Intermountain Healthcare; 5Primary Children Medical Center, Division of Inpatient Medicine, Salt Lake City, Utah

Introduction
Acute Respiratory Distress Syndrome (ARDS) is a severe form of respiratory failure that impairs normal lung function resulting in deficient oxygenation leading to increased morbidity and mortality among critically-ill patients. Mortality rates range from 26% - 58% and increase based on severity of disease. Incidence among ICU patients is reported as high as 10.4% and increases to 23.4% in patients on mechanical ventilation. The current diagnostic standard for ARDS is the Berlin definition which has three components: 1) acute onset within 1 week of clinical insult or new/worsening respiratory symptoms; 2) PaO2/FiO2 ≤ 300; and 3) radiographic evidence of bilateral infiltrates. Early and accurate identification of ARDS may lead to more expeditious treatment with lung protective ventilation that can result in positive outcomes for critically-ill patients that are already battling with insufficient life-sustaining reserves. To address this need for early and accurate identification, we are in the process of developing a real-time ARDS case detection system that will automatically identify ICU patients at high risk for ARDS. The major challenge and critical component of this system is a natural language processing algorithm that can interpret radiograph reports for bilateral infiltrates consistent with the Berlin definition with a high level of accuracy. We describe our retrospective results demonstrating the efficacy of using natural language processing to identify radiograph reports consistent with the presence of bilateral infiltrates.

Methods
A reference set was constructed from previously confirmed ARDS cases (n = 1,171) included in several NIH sponsored ARDSNet studies where Intermountain Healthcare was a participating study site. This set was augmented with negative cases (n = 3,513) for whom ARDS was a possible diagnosis but was ruled-out as not meeting the Berlin definition for ARDS. Cases ranged over the period from August 11th, 2000 to March 12th, 2018 and were cared for at an Intermountain facility in Salt Lake City, Utah. One-view and two-view x-ray studies were pulled from Intermountain’s Enterprise Data Warehouse. A radiograph study range was defined as all films 3 hours prior to intubation and 72 hours post intubation date. Study inclusion criteria for NLP processing was set as 1) interpret all studies 24 hours post intubation; 2) if no studies were found 24 hours post intubation, include all films 3 hours prior to intubation; 3) if still no studies are found, interpret the first study found between 48 and 72 hours post intubation. For cases that did not have an intubation date index, studies were pulled within a 72 hours window post admit date index. Any positive interpretation for bilateral infiltrates or direct mentions indicative of ARDS among the set of interpreted reports was considered positive overall for ARDS. Several machine learning models including SVMs, Random Forest, and Gradient Boosted Trees were considered for document level classification using an n-gram feature space. Performance was assessed using 10-fold cross validation, training folds were down sampled to balance positive/negative cases while test folds maintained case prevalence.

Results
Controlling for over-fitting, the Random Forest (401 trees) was the best performing model. Sensitivity: 0.88, Specificity: 0.80, Positive Predictive Value: 0.81, Negative Predictive Value: 0.87, Accuracy: 0.84, AUC: 0.84.

Conclusion
Natural language processing can be an effective clinical decision support tool to assist in the real-time identification of ARDS cases. Early identification may lead to early effective treatment, positively impacting outcomes of this life-threatening disease. Further work is underway to optimize NLP operating characteristics by using a more sophisticated phrase level entity recognition strategy that further disambiguates complex diagnostic mentions and harnesses partial evidence towards ARDS by preserving clinical context more effectively.
Exploring Marijuana Use in Patients Suffering from Pain

Dezon Finch, PhD, Stephen Luther, PhD
1James A Haley Veterans Hospital, Tampa FL

Introduction

While marijuana is by federal law an illegal substance and is considered an illicit drug by the Veterans Health Administration there is a movement to evaluate its use for therapeutic purposes1. Our objective is to determine if information on the use of marijuana is recorded in medical documents such that it could be reliably extracted with the use of Natural Language Processing.

Methods

As a part of a jointly funded (HSR&D and the NIH) retrospective cohort observational study to develop pain care quality indicators, NLP was used to extract evidence of marijuana use from medical progress notes. A corpus of 4400 medical progress notes from patients who suffer from pain in the primary care setting was annotated by two clinicians and adjudicated by a third expert to develop NLP. One of the concepts targeted was the use of Complimentary and Integrative Health modalities which included the use of marijuana. A list of terms representing marijuana use was developed from the annotations and supplemented from other sources. A simple regular expression-based extraction method was developed on the annotated reference set (F-Measure = .89) and applied to a corpus of 152,620 medical progress notes.2.

Results

We found a total of 12,518 separate mentions of marijuana in the corpus. We found marijuana mentioned in a variety of ways. Marijuana use was mentioned in cases where the patient was found positive for THC, was a past user, current user, in rehab, and as a user for medical purposes requiring a state card. In many cases notations were made that the Veteran was using marijuana to reduce pain or for reducing anxiety. Claims were often documented that the use of marijuana was effective, in some cases “the only thing that works.”

Conclusion

While this effort is largely exploratory, our efforts show that evidence of marijuana use is documented in medical documents in a way that can be reliably extracted and is talked about in a variety of circumstances. In some cases it appears to be useful for the treatment of pain or anxiety. Studies on the prevalence of marijuana use among Veterans who suffer from pain could be determined as well as the reason for its use and its effectiveness as a replacement for opioids. For Veterans who suffer from chronic pain: Is there a lower rate of opioid use among Veterans who use marijuana than those who do not use it?

References


1574
Generating Synthetic Control Subjects Using Machine Learning for Alzheimer's Disease Clinical Trials

Charles K. Fisher, PhD, Yannick Pouliot, PhD, Aaron M. Smith, Jonathan R. Walsh, PhD
Unlearn.AI, San Francisco, CA

Introduction

Alzheimer’s disease (AD) is a complex neurodegenerative disease with multiple cognitive and behavioral symptoms. Since recruitment of subjects for clinical trials in AD is challenging, we propose a machine learning approach to address this problem. Specifically, we use data from control subjects of past clinical trials to train an unsupervised machine learning algorithm to model the clinical data of control subjects of future trials. We term these machine learning-generated control subjects “synthetic control subjects.”

Method

We developed an unsupervised model of AD progression using a Conditional Restricted Boltzmann Machine (CRBM)\(^1\). We trained this model using data from 1,908 control subjects from 28 randomized clinical trials (RCTs) involving mild to moderate AD, provided by the Coalition Against Major Diseases Online Data Repository for AD\(^2\). Specifically, we trained the model on 44 variables measured across 18 months in 3-month intervals.

Results

We validated our model by comparing synthetic control subjects generated from the CRBM to actual control subjects not involved in training the model. The synthetic control subjects were statistically indistinguishable from actual control subjects across multiple measures, with only minor differences identified. The Alzheimer’s Disease Assessment Scale–Cognitive Subscale (ADAS-Cog\(^3\)) is a prominent multi-component cognitive assessment scale used frequently as an endpoint in AD RCTs. We further validated our model by predicting the ADAS-Cog progression of control subjects in a past RCT, finding good agreement with the observed progression of the actual control subjects. We then compared CRBM-generated ADAS-Cog scores to those generated by supervised neural network, random forest, and linear regression models trained on the same data. All algorithms exhibited a similar error profile, with CRBM generally exhibiting the best performance. These comparisons provide evidence that our unsupervised learning model can generate accurate synthetic control subjects.

Conclusion

This work demonstrates the potential for the model described here to generate synthetic control subjects that may be used to supplement or replace actual control subjects and reduce the reliance on actual control subjects in randomized clinical trials.

References

Prescription opioid tapering: Mining structured and unstructured data

Samah J. Fodeh\textsuperscript{a,b}, E. Jennifer Edelman\textsuperscript{a}, William Becker\textsuperscript{a,b}, Cynthia Brandt\textsuperscript{a,b}, Sally Haskell\textsuperscript{a,b}

\textsuperscript{a} Yale School of Medicine, New Haven, CT, USA
\textsuperscript{b} VA Connecticut Healthcare System, West Haven, CT, USA

Introduction: In response to significant evidence of harms associated with prescribed opioids (e.g., opioid misuse, opioid use disorder and overdose deaths), the Centers for Disease Control and Prevention (CDC) issued the “CDC Guideline for Prescribing Opioids for Chronic Pain” in 2016.\textsuperscript{[1]} In this guideline, it is recommended that when “benefits do not outweigh harms of continued opioid therapy, clinicians should optimize other therapies and work with patients to taper opioids to lower dosages or to taper and discontinue opioids.” In particular, opioid taper to a reduced dosage or tapering and discontinuing opioid therapy is encouraged in specific contexts, such as patient requested dose reduction; lack of meaningful improvement in pain and function; and prescribed dose of $>50$ morphine milligram equivalents per day without benefit or combined opioid/benzodiazepine therapy. Given lack of robust evidence to guide tapering strategies, a dose reduction occurring at 10\% of the original dose per week is generally recommended with slower taper if symptoms arise and quicker tapers in the context of serious adverse events. These recommendations are consistent with other United States-based and international guidelines. Currently, the field lacks accurate and reliable approaches to identifying patients undergoing tapering using structured pharmacy fill/refill data which further limits our ability to understand the impact of tapering. In addition to structured data, we explored the utility of using text narratives in clinical notes to detect patients being tapered.

Methods: We identified patients on Long Term Opioid Therapy (LTOT) using pharmacy data (based on CN101 class of medications). LTOT is defined as receipt of more than one opioid prescription with 90 days supply within a given year allowing 30 days gap in between fills. Using the Women Veterans Cohort Study (WVCS) at the Veterans Healthcare system, we used Voogo\textsuperscript{[2]}, a search engine that supports both free text and structured data searches, to retrieve clinical notes in FY (2015-2017) on LTOT patients. We explored tapering mentions in primary care notes and substance abuse notes in FY2015 and FY2016 only for lack of access to FY2017 clinical notes. We ran multiple queries using a list of search terms including: taper, discontinue, lower dose and their combinations with medication names (e.g. taper methadone, discontinue oxycodone) and allowing for a word or two in the context.

Results: Using pharmacy data, we identified 36,365 patients receiving LTOT with 83,457 opioid prescriptions between FY2015 and FY2017. We collected 441,106 and 712,215 clinical notes in the FYs 2015 and 2016, respectively. A slight decrease in the number of patients and prescriptions between FY2015 and FY2016 is noticed compared to a more dramatic decrease from FY2016 to FY2017. Table 1 shows that 854 (5\% of LTOT patients in FY2015) have evidence of opioid taper discussions in their clinical notes compared to 456 LTOT patients (3\%) in FY2016. Only 25\% of these discussions occurred during their visits to specialty clinics as reflected by the clinical note type (substance use disorder) and the remaining 75\% occurred during their primary care visits. In FY2016, 3\% of LTOT patients had opioid taper discussions in their notes, 63\% of which happen at primary care.

Discussion: Pharmacy data reflected adherence to prescription guidelines as published by the CDC late 2016 (Figure 1). Using unstructured clinical notes, we had the capacity to identify LTOT patients who are being tapered. Searching FY2015 clinical notes for combinations of the terms taper and each opioid medication, we retrieved 323 notes as a result of the query. By adding the terms discontinue and lower dose, and substance disorder we identified 221, 519 and 391 more, respectively. In a similar manner, we identified 108, 97, 220 and 343 notes in FY2016, in the same respective order. This finding highlights the importance of mining clinical notes to help characterize the different tapering strategies not captured using pharmacy data. Using pharmacy data, we identified a high proportion of patients with prescription opioid receipt with patterns likely reflecting changes in opioid regulations (i.e., changes in hydrocodone scheduling) and guidelines recommending more judicious opioid prescribing. Using unstructured data, we captured patients being tapered with evidence of tapering discussions in clinical notes. We plan to enhance our approach by refining our definitions of long-term opioid receipt and applying more advanced computational approaches to not only detect tapered patients but also characterize the tapering strategies by different clinicians and how they affect clinical outcomes.

Acknowledgement: Pain, Research, Informatics, Medical comorbidities, Education center at the VA who funded this work.

References
Immune-Related Adverse Events of Checkpoint Inhibitors

Paul Fontelo, MD, MPH; Fang Liu, MS
National Library of Medicine, Bethesda, Maryland

Background

The immunotherapy of cancers, particularly checkpoint point inhibitors (CIs), have revolutionized the treatment of malignancies leading to increasing survival. They have been mostly beneficial in patients with melanoma and certain types of lung tumors. Cancer cells evade detection and destruction by the body’s immune system by acquiring a coat of normal proteins on their surface. CIs block these protective “normal” proteins and allows them to be recognized as “foreign” cells and destroyed by the immune system. They also inhibit the proteins on T-cells that recognize normal cells and allows functioning T-cells to destroy malignant cells. However, because they are immune inhibitors, they may also trigger T-cells to attack normal cells which result in Immune-related Adverse Events (IrAEs). These side effects may cause serious harm, and death, even after the drug is discontinued. CIs are increasingly used in treating cancers unresponsive or resistant to chemotherapeutic agents or even as a primary line of treatment of malignancies. Reports of IrAEs, especially when used in combination, have increased with their growing usage.

Design

We reviewed publications in PubMed of IrAEs for seven CIs and 21 two-drug combinations from 2005 to 2017. The search strategy used was as follows: drug name(s) + organ or system or condition + human + hasabstract. A search tool was developed for finding reports of IrAEs (https://go.usa.gov/xRmHK).

Results

Immune-related adverse events report started appearing in the literature in 2005 (Ipilimumab) but the majority of reports were published between 2014 to 2016 when more CIs were approved by the FDA. Most adverse events reported were dermatologic, followed by endocrine, gastrointestinal, respiratory and neurologic. Single drug IrAEs are shown in Figure 1. For combined regimens, pembrolizumab and Nivolumab combination accounted for 25% of all reported adverse events (2888) in the literature, followed by nivolumab AND ipilimumab (22%), then pembrolizumab AND ipilimumab (16%). Other drug combinations were each less than 10% of the total. Among organ systems, dermatologic (24%), endocrine (23%) and gastrointestinal (16%), respiratory (15%) and neurologic (13%) side effects were the most common adverse reactions to checkpoint inhibitors. For two-drug combinations, Pembrolizumab and Nivolumab (25%), Nivolumab and Ipilimumab (22%) and Pembrolizumab and Ipilimumab (16%) account for majority all immune-related adverse events (2888) reported.

![# of Publications in PubMed by Year](chart.png)

Fig. 1. Number of Immune-related adverse events publications for one-drug regimens by year

Conclusion

The incidence of IrAEs to CIs has grown significantly with their increasing use. Dermatologic, endocrine and gastrointestinal are the commonly affected. For single drug regimens, ipilimumab, nivolumab and pembrolizumab account for the majority of immune side effects. Among the two-drug combinations, Pembrolizumab and Nivolumab, Nivolumab and Ipilimumab and Pembrolizumab and Ipilimumab account for 58% of all adverse events.
mHealth for Weight Management Among Breast Cancer Survivors from Vulnerable Populations

Marva Foster, PhD, RN, Department of Veterans Affairs, Boston, MA; Lauren J. Oshry, MD, Boston University, Department of Medicine, Section of Hematology and Medical Oncology, Boston, MA; Lisa M. Quintiliani, PhD, Boston University, Department of Medicine, Section of General Internal Medicine, Boston, MA

Introduction

Advances in detection and treatment of breast cancer has led to a large cohort of cancer survivors who may have an increased awareness of health promotion; mHealth technologies have the potential to deliver effective behavioral interventions in a format suited to everyday life. While researchers have identified Behavior Change Techniques (BCTs; Taxonomy version 1) present in efficacious interventions for weight loss among cancer survivors, the features that should be included in mHealth interventions for cancer survivors have not been well specified, particularly among those from low income or racial/ethnic minority groups.(2,3) Purpose: The objective of this study is to examine perceptions of evidenced-based features that breast cancer survivors from underserved populations would find important to include in a mHealth weight management app. Method: Female breast cancer survivors were recruited from a large safety net hospital to participate in an individual interview or small group discussion. This study was approved by the Boston University Medical Campus/Boston Medical Center IRB. In phase 1, participants were shown a presentation depicting a BCT, one at a time, along with accompanying pictures depicting how each BCT could be incorporated into a mHealth app. Participants rated how important each feature would be from 1 (not at all important) to 9 (very important). Using a modified Delphi method, each participant was then asked to re-rate the features after seeing their own rating compared to the group rating. Participants also answered open-ended questions about their perceptions of mHealth apps. Evaluation results: In phase 1, thirteen women participated; the majority were from a racial/ethnic minority group (85%), and received medicaid (65%). All were overweight or obese. Upon re-rating after seeing the group average, scores ranged from 6.1 -8.6 for the BCTs. Lower rated BCTs included the need for specifying credible sources; higher rated features included feedback and self-monitoring of outcome of behavior (weight). Participants valued mHealth apps that were simple to use “It has to be simple enough…I’m not a technical person. So, it has to be simple and somebody like me can use [it]”; had new/engaging information, “…as long as it was interesting reading, if that makes any sense. ... If it’s more funny, or quick, or easy to read, interesting and incorporated some features specifically for breast cancer survivors. In phase 2, 9 women participated. The majority (n=8) rated the usability of the mock-ups in the good/excellent range (scores ≥68). Conclusions: Breast cancer survivors from underserved populations were able to identify important features to include in a mHealth app and to elaborate on preferred app design characteristics. Future research will involve further refinement of the mHealth app, ultimately resulting in a mHealth app ready for efficacy testing in a larger trial.

Planned modifications to the study and app

1. Perform field testing of the app, make further refinements based on user feedback.
2. Test app in a larger efficacy trial.

References

SNOMED CT: Interoperable but silos remain between medicine and nursing

Dan Fraczkowski¹, RN, Andrew D Boyd ², MD, Karen Dunn Lopez³, PhD
¹ College of Nursing, University of Illinois at Chicago- Chicago, IL, USA ² College of Applied Health Sciences, University of Illinois at Chicago- Chicago, IL, USA

Introduction

Transitions of care can impair communication and lead to adverse events. Our prior work examined the lack common terms between the professions caring for the same patient. [1] SNOMED CT is an interoperable healthcare terminology with potential to facilitate interprofessional communication. [2] NANDA-I[3], more commonly known as nursing diagnoses, contained in SNOMED CT, represents the nursing diagnostic judgements. The objective of this study was to identify patterns of terms used by physicians in their discharge summaries that map to NANDA-I terms/concepts.

Methods

We performed a secondary data analysis merging 3 data sets: 1. Physician discharge summary Concept Unique Identifiers (CUIs) from a large urban tertiary care hospital for a six year period. 2006-2012 (12,968 terms). CUIs were assigned using MEDlee across all disciplines and over 100,000 notes. [1] CUI’s were used at least 15 times/year. 2. SNOMED CT US nursing terms(673 terms)[4]. 3. NANDA-I list of nursing diagnosis (244 terms)[5]. CUI terms were identified via manually searching in the UMLS and compared against the SNOMED CT database

Results

Of the physician CUI terms, we identified 159 (23%) terms in nursing subset within discharge summaries. The following NANDA-I categories were represented: 87 defining characteristics, 45 associated conditions, 38 related factors, 27 risk factors and 17 population at risk for (CUI could belong to several). Among the SNOMED CT Nursing terms, 21%(33) unique CUI’s mapped directly to NANDA-I. These encompass 6 of the 13 NANDA-I domains, absent are health promotion, life principles, growth and development, coping/stress tolerance and role relationship.

Discussion

Our results show that despite continued development of SNOMED CT with greater inclusion of nursing terms, from 368 to 673 in 2017 [3] physicians do use some nursing concerns in discharge summaries. However, the minimal presence of nursing documentation in physician summaries is a limitation of SNOMED CT as a universal terminology. Also concerning in terms of semantic interoperability, physician terms that map to the SNOMED CT nursing subset leave room for ambiguity, for example, “hyperthermia” has 7 associated conditions, 11 defining characteristics, 2 risk factors and 1 related factor. The associated conditions varied from, “Sepsis” to “Increase in metabolic rate.” Finally, the omissions of the six domains reflect care that may be a unique focus for nurses, reflecting gaps and opportunities to improve the patient care record.

Conclusion

Standardized terminologies are proposed as a means to “bridge the gap” however nursing and physician care are unique. Our results indicate little use of nursing terms in physician discharge summaries, thus we haven’t yet reached full integration with these terminologies in practice. The limited level of crossover indicates challenges remain as we seek for SNOMED CT to reach full potential.

References

Evaluating the performance of a general diagnostic decision support system on case data from urgent cardiology admissions

Fraser HSF, MBChB MSc
Brown University, Providence, RI, USA

Abstract

Many new diagnostic decision support systems have been introduced in recent years but lack rigorous clinical evaluation or even use of real patient data. We used data on urgent cardiology admissions to test an established general diagnosis program Isabel. Results were similar to a specialised Heart disease diagnosis program (HDP).

Introduction:

Misdiagnosis is a common problem affecting 5% of ambulatory care patients in the US. It can lead to poor investigation and management of patients, and put lives at risk. Diagnostic decision support systems go back to the beginning of medical informatics, and underwent a resurgence over the last 5 years. This likely stems from better access to medical data, better algorithms, and better delivery methods through EHRs or mobile devices. However there is a lack of rigorous evaluation using real patient presentations and validated clinical diagnoses rather than the use of case scenarios created by physicians. This study used case data on urgent admissions to a cardiology ward entered in real time by medical residents with “gold standard” diagnoses from detailed chart review and follow up, to test “Isabel” a well-established general medical diagnosis program.

Methods:

Data on urgent admissions to the cardiology service at the New England Medical Center in Boston, MA was entered into a web based clinical interface by the admitting residents. The study tested the diagnostic performance of the Heart Disease Program (HDP) 1 which was built with a Bayesian belief network modified to handle temporal and severity characteristics of clinical data. After the data was entered the residents were required to enter their own differential diagnosis before they could view the HDP diagnosis. The patients were then followed up after their admission to determine the final clinical diagnoses. In the current study the diagnostic performance of the general diagnosis program Isabel 2 was tested using the clinical data entered into the HDP in the earlier study. The differential diagnosis from Isabel was then coded with the same vocabulary as the HDP. Diagnostic performance of Isabel against the gold standard diagnoses was compared with the performance of the resident physicians and the HDP also against the gold standard. Performance metrics were Comprehensiveness (similar to sensitivity), and Relevance (similar to Positive Predictive value), calculated per case and reported as mean values.

Results:

Isabel was tested with 16 sequential cases from the original study including demographics, clinical history, cardiac risk factors, physical exam findings and investigations. Diagnoses included hypertension, myocardial infarction, angina, valvular heart disease and congestive heart failure. The Comprehensiveness of Isabel was 56% compared to 46% for the residents and 61% for the HDP. The Relevance was 19% for Isabel versus 51% for the residents and 26% for the HDP, the lower scores for the diagnosis programs reflected larger numbers of diagnoses per case.

Discussion and conclusions:

Comprehensiveness of 50 – 60% is not surprising given that the gold standard was based on follow up and extensive further investigation. Data entry into Isabel was significantly less detailed than the HDP, particularly timing of symptoms, numeric data for vital signs, cardiac exam and investigations. This may have contributed to the 5% lower Comprehensiveness of Isabel, which also covers a much more complete set of adult and paediatric diagnoses. This is part of a larger study of diagnostic programs using real patient data entered in real time.

References

Incorporating Social Determinants of Health in Electronic Health Records (EHRs): A Qualitative Study of Perspectives among EHR Vendors

Maysoun Freij,1 Prashila Dullabh,1 Sarah Lewis,2 Scott Smith,3 Lauren Hovey,1 Rina Dhopeshwarkar1

1NORC at the University of Chicago, Bethesda, MD; 2 Centers for Medicare and Medicaid Innovation, Baltimore, MD; 3Assistant Secretary for Planning and Evaluation, Washington D.C.

Introduction

Health care reform initiatives over the past decade have simultaneously incentivized value-based care payment models and the adoption and development of electronic health records (EHRs).1 Further, the emphasis on value over volume has drawn attention to the importance of social determinants of health (SDH) (i.e., socioeconomic issues such as housing, food, transportation, income, education, social isolation) in potentially affecting health outcomes.

Methods

We conducted a literature review on current uses of EHRs for assessing and documenting SDH. To further investigate the development of SDH-related software products, we conducted interviews with EHRs vendor representatives purposefully selected based on vendor market shares in both ambulatory and inpatient settings. We recruited six vendors with large market shares in one or both settings. Interviews were conducted by phone, and notes were transcribed. We qualitatively analyzed interview notes using NVIVO software for (1) motivations to develop such software products (2) descriptions of their products and uses, and (3) facilitators and challenges to collection and use of SDH data. This study was approved by NORC’s Institutional Review Board; all participants gave informed consent, and agreed to the confidential presentation of their findings.

Findings

Our findings indicate that vendor systems and their functionalities are influenced by client demand and initiative, federal initiatives, and the vendors’ strategic vision about opportunities in the health care system. Among the sampled vendors, SDH is a new area for growth, and the vendors range in the number and sophistication of their SDH-related products. While all participating vendors have enabled some SDH data collection screening instruments or measures in their EHR platform, they varied in terms of capacity to track referrals and analyze data. Vendor representatives reported a range of activities within this field, ranging from helping clients meet regulatory obligations to developing SDH-related products for use in population health management. They identified a number of challenges with analyzing SDH data and sharing them among health systems, including multiple overlapping but distinct performance metrics and indicators across various federal and state programs, lack of agreement on mapping SDH measures to codes, and lack of codes for all measures. Vendor representatives recognized the need for more standardization of SDH performance measures across various federal and state programs, better mapping of SDH measures to multiple types of codes, and development of more codes for all SDH measures of interest.

Conclusion

Though EHR vendors have both indirect and direct roles in working with policymakers and health care systems on value-based care and health care quality initiatives, there are few systematic studies of the EHR vendors and their motivations. Their perspectives are important given that product-specific decisions may end up becoming de-facto policies given the market shares of particular vendors, particularly in areas where policy guidance is lacking, as it is for SDH data collection and use. Commercial vendors appear ready to collaboratively discuss policy solutions, such as standards or guidelines with each other, health care systems, and government agencies in order to further promote integration of SDH data into the standard of care for all health systems. Findings from this study have been published by the Office of the Assistant Secretary for Planning and Evaluation, and will appear in a forthcoming report to Congress.

References

Conversion of a Complex Legacy Antimicrobial Stewardship Clinical Decision Support System into Epic

Brian K. Fung, PharmD1, Evan Draper, PharmD1, James Steckelberg, MD1, Abinash Virk, MD, Sara Ausman, PharmD1, Sarah Lessard, PharmD1, Aaron Tande, MD1, Lynn L. Estes, PharmD1
1Mayo Clinic, Rochester, MN, 2Mayo Clinic Health System, Eau Claire, WI, 3Mayo Clinic Health System, La Crosse, WI

Background

The hospital rules-based system (HRBS) was the first rules-based informatics program developed at Mayo Clinic in the 1980s in an effort to improve patient care through the reduction of medication errors, enhanced education, and clinician communication. Computer-based antimicrobial monitoring (CBAM) was the first of the six HRBS subsystems to be developed and demonstrated positive patient outcomes and favorable cost reductions in medications. In January 2015, Mayo Clinic announced the selection of Epic as its partner for a single, integrated electronic health record (EHR). We describe the conversion of Mayo Clinic’s legacy decision support system CBAM into Epic’s antimicrobial stewardship program (ASP) module for eighty hospitals and clinics across four states.

Methods

The project kicked off in July 2015 and planning began between project analysts and subject matter experts. Epic’s antimicrobial stewardship strategy handbook was used to guide core implementation and build principles. Due to insufficient functionality using Epic’s foundation tools, in April 2016, the strategy shifted from using foundation Epic build to designing a customized ASP module that mirrored the complexity of Mayo’s CBAM system.

Results

The design, build, and implementation required substantial effort from key stakeholders from infectious disease physicians, infectious disease pharmacists, and project analysts. Bi-monthly meetings were set up in March 2016 to facilitate convergence and rapidly mitigate issues as it pertained to the Epic build. Over 200 legacy antimicrobial stewardship rules were reviewed, converged, and replicated into 75 customized rules for build in Epic. Estimated effort for rule build varied between 5 hours to 30 hours. Of the planned 75 rules, 17 were unable to be replicated by the first go-live in July 2017 due to complexity and constraints of the existing EHR tools. Project analysts innovated a solution that combined multiple tools in nontraditional methods to replicate minimum functionality that included rule interaction (i.e. dismiss and defer) and rule complexity (e.g. relative properties). A prototype of two rules using the new framework was introduced in the second go-live in November 2017. Given the successful implementation of the prototype, a full-scale conversion of existing rules and completion of the remaining 17 rules were implemented by the third go-live in May 2018. While this go-live marked the completion of converting the legacy systems into Epic, the initial weeks were characterized by false positives and usability issues. Significant optimization efforts of the ASP module persisted until the fourth go-live in October 2018 and the project officially closed on January 1, 2019. Despite the challenges, the ASP module is markedly improved from its earlier prototypes and its success and progression has served to inform Epic on multiple areas for future research and development.

Discussion

Implementation of Mayo Clinic’s legacy decision support systems for antimicrobial stewardship was achieved within project timelines. Innovative solutions with existing tools were required to fully replicate legacy functionality that allowed for interaction with individual rules.

References

Corpus Size Influences Clinical Concept Embeddings

Cheng Gao PhD\textsuperscript{1}, Chao Yan MS\textsuperscript{2}, Bradley A. Malin PhD\textsuperscript{1,2}, You Chen PhD\textsuperscript{1}
\textsuperscript{1}Vanderbilt University Medical Center; \textsuperscript{2}Vanderbilt University

Introduction

Word embedding has shown great promise for many natural language processing tasks. It is a general belief that a larger corpus results in better word embedding. However, a recent work\textsuperscript{1} showed that the combination of PubMed and PMC on clinical text generated worse embeddings than PubMed alone, indicating that a larger corpus cannot guarantee better word representation. In this study, we investigated in learning of International Classification of Disease, 9th Revision (ICD-9) codes how the corpus size influences word embedding. Specifically, we used word2vec\textsuperscript{2} to train word vectors of ICD-9 codes from de-identified data and fed these word vectors to deep learning model to predict neonatal encephalopathy (NE). We found that the word embeddings learnt from a smaller corpus had a better performance in predicting NE comparing to a larger one.

Materials and Methods

This study relies on two corpora of timestamped ICD-9 codes to learn word vectors: i) Global dataset: all Vanderbilt University Medical Center (VUMC) patients – approximately two million; ii) Local dataset: 31,158 maternal patients at VUMC. Word2vec is applied to learn word embeddings from these two datasets. To investigate which dataset generates better embeddings for ICD-9 codes, we evaluate these embeddings intrinsically and extrinsically.

Intrinsic Evaluation: CCS, including 285 clinically meaningful categories, is used as a gold standard for ICD-9 codes. For each code, we perform nearest neighbor (NN) search based on the word vectors (word2vec\textsuperscript{2}). If the code and its NN belong to the same CCS category, we consider the word vector for this code is correct. The correct rate will be reported along with 95% confidence interval.

Extrinsic Evaluation: We feed learned word vectors from Global and Local datasets as inputs into long short-term memory (LSTM) network to predict NE. This is a particularly challenging task due to the heavily unbalanced nature of its classes: 104 NE cases and 31,055 patients. We use 10-fold cross-validation and report the area under the receiver operating characteristic curve (AUC) with a 95% confidence interval on test data.

Results and Discussion

Table 1 summarizes the word embedding performance for both datasets, with the best results in bold font. It can be seen that the embeddings learnt from Global dataset are better when they are intrinsically evaluated, implying that a larger corpus generates better embeddings in identifying similar ICD-9 codes. However, the local dataset yielded the best extrinsic results in predicting NE and the statistical test (Delong method) show that AUCs are significantly different with p value of 0.005. This implies that locally-learned word embeddings are more appropriate for the NE prediction task. Since NE prediction is the only task tested in this study, the result may be different in other clinical datasets. It is suggested comparing AUC results from global dataset and local dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Proportion to Learn Word Vectors</th>
<th>Intrinsic Evaluation (CSS) Percent (95% confidence interval)</th>
<th>AUC (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global dataset</td>
<td>100%</td>
<td>0.815 (0.808 – 0.822)</td>
<td>0.899 (0.875 – 0.922)</td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td>0.529 (0.521 – 0.537)</td>
<td>0.881 (0.848 – 0.914)</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>0.515 (0.505 – 0.525)</td>
<td>0.832 (0.734 – 0.931)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>0.483 (0.477 – 0.489)</td>
<td>0.815 (0.763 – 0.868)</td>
</tr>
<tr>
<td>Local dataset</td>
<td>100%</td>
<td>0.539 (0.531 – 0.546)</td>
<td>0.933 (0.917 – 0.949)</td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td>0.506 (0.501 – 0.511)</td>
<td>0.928 (0.904 – 0.952)</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>0.492 (0.486 – 0.498)</td>
<td>0.920 (0.900 – 0.939)</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>0.489 (0.484 – 0.494)</td>
<td>0.919 (0.898 – 0.940)</td>
</tr>
</tbody>
</table>

References

Application of IBM Watson to clinical decision support: a preliminary experience on the interpretation of DSM-5

Enrico Garbelotto, MS¹, Marco Prenassi¹, PhD, Sara Marceglia, PhD¹
¹Dipartimento di Ingegneria e Architettura, Università degli Studi di Trieste, Trieste, Italy

Introduction
The Diagnostic and Statistical Manual of Mental Disorders launched its 5th edition (DSM-5) in 2013¹. The manual is a reference for psychiatric and neuropsychiatric practice and introduced novel advances in disorders diagnosis and treatment thanks to the recent research in the field. Despite the attempts to decrease the total number of categories, the manual remains very complex, and includes a large co-occurrence of diagnoses (sometimes known as comorbidities), with sometimes a lack of distinctiveness between diagnoses². This work aimed to test whether a system based on IBM Watson services could be implemented and used to support the clinician in an immediate and easy consultation of the DSM-5, also simplifying the identification of the most probable diagnoses.

Methods
Three IBM Watson services, namely Knowledge Studio, Discovery, and Natural Language Understanding (NLU) were used to build up the system for DSM-5 interpretation. A concept model was developed and trained in Watson Knowledge Studio using 60 extracts from 6 DSM-5 chapters previously annotated according to the entities and relationships defined in the concept model. The precision of the system (percentage of true positives on the overall positives recognized by the model) and the recall (percentage of true positives on the overall true from the ground truth) were used as performance measures of the model. A Java-based application was then implemented as user interface for the two other IBM Watson services (Discovery and NLU)³. The system was validated retrospectively using 5 case studies available in the literature, in order to test its ability to understand the clinical observations and to suggest the possible diagnosis/es according to the DSM-5.

Results
The semantic model included 9 main entities (among them, “disorder”, “subtype”, “specifier”, “diagnostic criterion”, “risk factor”) and 9 main relationships and achieved an overall precision of 85% and recall of 76%.

The application implements three main functions: (1) to retrieve the parts of the DSM-5 relevant for explaining specific clinical observations; (2) to obtain a list of probable diagnoses (with associated likelihood) starting from a set of clinical observations (diagnostic criteria, risk factors, or specifiers) listed in natural language; and (3) to extract the relevant concepts, according to the semantic model, provided an extract of the DSM-5. Whereas the first two functions can be used directly by clinicians, the third one is more useful for implementing other applications.

Examining 5 retrospective cases, we found that in all cases the correct diagnosis was reported in the list of probable diagnoses, being the first one in 4 cases (Panic Disorder, Post-traumatic Stress Disorder, Illness Anxiety Disorder, and Generalized Anxiety Disorder), and the second one in one case (Borderline Personality Disorder).

Conclusion
Our results show that IBM Watson services can be used to support the reading and interpretation of the DSM-5, and the diagnostic process. Even in a first implementation and testing, the system correctly recognized all the clinical observations and possible associated diagnoses. However, the validation should include a real-world testing and an improvement of the semantic model.

References
Factorial Design Survey Methodology on REDCap and Qualtrics: A Comparative Analysis

Jose P. Garcia, Jr., BA1, Sarah Collins Rossetti, RN, PhD3,4, Kenrick D. Cato, RN, PhD4, Suzanne Bakken, RN, PhD, FAAN, FACMI3, Haomiao Jia, PhD4, Min J. Kang, RN, PhD1,2, Christopher Knaplund, MPhil4, Patricia C. Dykes, RN, PhD, FAAN, FACMI1,2

1Brigham & Women’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Columbia University, Department of Biomedical Informatics, New York, NY; 4Columbia University, School of Nursing, New York, NY

Problem

A factorial design survey (FDS) is an experimental design that includes multiple iterations of a survey, where the factor content and order change from iteration to iteration to determine which combination of statistically significant factors are perceived as clinically significant, yielding more generalizable conclusions. The beauty of FDS is that it combines the strengths of a factorial experimental research design with a sample survey, but it can employ ten or more variables to study within one survey.1 An appropriate FDS software is needed to reduce the complexity associated with measuring the effect of each variable over numerous trials in both survey development and end-user experience to ensure design worktime is feasible and an acceptable response rate is achieved. This case study compares the ability of two popular survey platforms (REDCap and Qualtrics) to efficiently run an FDS trial.

Methods

A pilot survey trial was conducted on both REDCap and Qualtrics. REDCap lacks on-platform survey field randomization, so a workaround was developed. An off-platform block randomizer was used to generate and randomly assign surveys containing four vignettes to participants, and a template survey containing fields with every possible vignette (n=16) hidden from survey view was created on REDCap. To create test surveys, copies of the template were made, and the fields containing the vignettes assigned by the block randomizer were unhidden, leaving only the vignettes (n=4) randomly assigned to each participant in their survey view. The pilot survey on Qualtrics was conducted completely on platform. A master survey was created using Qualtrics’ Randomizer element in Survey Flow that would randomly populate embedded fields with one of two levels for each factor (n=5) in every test vignette (n=6) for each survey recipient. The surveys for each trial were disseminated through their respective web-based tools. Time required to develop the surveys and end-user perceived ease of use were measured for both platforms. A student T-test was used to test perceived differences in ease of use.

Results

For REDCap, 42 surveys were created and disseminated in approximately three hours (~4.5 minutes/survey). One survey was created and sent to 34 participants in approximately one hour (~1.75 minutes/survey) on Qualtrics. The ease of use survey results for both software were not statistically different (p=0.19).

Conclusion

REDCap’s lack of randomization and automation functions are limitations for factorial design. The reliance on a workaround requiring manual input for executing FDS in REDCap introduces potential for human error when scaled up. Qualtrics, however, possesses the tools necessary to efficiently and automatically run a factorial design survey while showing no significant difference in participant-reported ease of use.

Acknowledgements

This study was funded by the National Institute of Nursing Research (NINR): 1R01NR016941-01, Communicating Narrative Concerns Entered by RNs (CONCERN): Clinical Decision Support Communication for Risky Patient States. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Querying Clinical Datasets with GraphQL and Concept IDs

Julie P. Garcia¹, Rimma Belenkaya, MS¹, Peter Stetson, MD, MA¹, and John Philip, MS¹
¹Memorial Sloan Kettering Cancer Center, New York, NY, USA

Introduction

At Memorial Sloan-Kettering, we are working on ways to automate the retrieval of patient data from clinical systems for research purposes. One approach is to combine a system for storing ontologies, such as TopBraid EDG, with a GraphQL Server to query data from many sources. GraphQL is a fast, flexible query language and server framework that allows you to combine multiple data sources in a single request.¹² This allows the aggregation of datasets across the organization with metadata (e.g. alternative labels, data provenance).

Using GraphQL to Query Data by Concept ID

We demonstrate a query schema that combines data with ontologies stored in TopBraid EDG. The GraphQL server allows for full-text searching to retrieve an associated concept ID and location of the data. For example, when searching for all concepts containing the term “specimen”, you would receive a list of concepts that contain this search term. A selected concept ID can then be used in combination with a patient identifier, such as a medical record number, to retrieve the actual data value or value set from the source. In this example, pathology specimen data from a Cerner CoPath Sybase Database is mapped with a concept ID stored in TopBraid (Figure 1a). The query asks for a concept value, given a concept ID and patient MRN (Figure 1b) and the data value or value set is returned.

![Figure 1](image.png)

Figure 1. a) GraphQL Server combines pathology data with concepts stored in TopBraid EDG to allow for single request from frontend platforms b) Query by Concept ID and patient MRN returns requested related data from associated source

This GraphQL API server can be used as the backend for various data analysis tools that include custom web applications, dashboard platforms and mobile apps. Multiple data sources can be queried at once, and only the data that is needed will be returned. This allows for rapid frontend development, with very little backend refactoring that is generally required with traditional RESTful APIs.²

Conclusion

Utilizing a GraphQL API server to aggregate data and metadata across an organization can simplify the process of retrieving data from multiple sources, while forcing use of standardized terminologies. This can rapidly accelerate research and reduce tedious manual data abstraction processes.

References

Description of Problem Addressed by Project:
Over the past six years many vaccine manufacturers have affixed two-dimensional (2D) barcodes onto vaccine vials and syringes; the majority of which are now 2D barcoded. Some electronic medical record (EMR) systems have incorporated 2D barcode functionality, applications to not only capture and process the barcoded data but to incorporate patient safety alerts. The Centers for Disease Control and Prevention (CDC) conducted several pilots to explore the impact of 2D vaccine barcode scanning on the data quality captured during vaccine administration. Healthcare practices considering the implementation of vaccine barcode scanning face not only technical challenges but also human considerations as staff adjust to and adopt new work flows and protocols.

Description of Project and Lessons Learned:
CDC has developed an implementation guide based on the experience with the pilots on how to scale the adoption of the 2D barcode scanning technology in the clinical setting. Specifically, this guide was developed as part of a pilot to implement 2D barcode scanning in a large healthcare system. Approximately half of the large healthcare system’s practice sites participated in the pilot. For a period of six months, participant sites used a 2D barcode scanner to capture vaccination data for patients receiving a vaccination. The pilot investigated when, why, and how barcode scanning should be implemented to achieve the highest return on investment, and how behavioral strategies can be used to improve consistent adoption over the long-term. This resulted in an implementation guide that is the basis for the poster.

Poster Content:
This poster will provide an overview of 2D vaccine barcoding in 2019 and summary of project lessons learned including how behavioral strategies (“nudges”) to promote scanning impacted scanning rates and longer-term adoption. The poster will also describe baseline scanning rates across groups, changes to rates based on implementation of behavioral strategies, resolution of key challenges, and insights about scalability within a single healthcare system. A detailed description of the steps to implement vaccine barcode scanning in clinical settings will feature prominently on the poster. This information is based on lessons learned in the actual implementation described above and involves six key steps: decide, plan, train, assess, adjust, and sustain. The poster will describe these steps and essential lessons that can be used when introducing a new technology, including how to design strategies to encourage staff to implement new protocols.
Usage Patterns of a SMART on FHIR Application to Improve Platelet Transfusion Workflow

William J Gordon MD MBI1,2,3, Maria Aguad BS1, Heather Appello BS1, Samuel J Aronson ALM2, Jane Baronas BA1, Indira Guleria PhD1,3, Richard M Kaufman MD1,3, Edgar Milford MD1,3, Michael Oates2, Rory dela Paz BS2, Melissa Yeung MD1,3, William J Lane MD PhD1,3

1Brigham and Women’s Hospital, Boston, MA, 2Partners Healthcare, Boston, MA 3Harvard Medical School, Boston, MA

Introduction

Worldwide, more than 50,000 patients undergo Hematopoietic Stem Cell Transplantation (HSCT) each year1, and many of these patients will develop transfusion-dependent thrombocytopenia. We have previously described2 a FHIR-based application designed to improve platelet utilization and transfusion workflow at a single center. We now report initial provider usage of this application along with patients loaded in the application.

Methods and Applications

We developed a visualization application to improve the workflow around platelet transfusion for patients undergoing HSCT. The application provides a unified view into several essential data elements relevant for patients requiring platelet transfusion, including platelet count trend, prior transfusions, patient HLA data, and current blood bank platelet inventory. This product was launched October 1st, 2018, at a single institution, after extensive validation and updates to the blood bank Standard Operating Procedures (SOPs). Initial usage of the application was high and has remained consistently high for the first 5 months of production use (Figure 1). Users include blood bank technicians, supervisors, physician trainees, and attendings. Usage has been limited to patients already identified as requiring Human Leukocyte Antigen (HLA) matching for platelets (for example, due to high calculated panel reactive antibody scores), though we are in the process of updating our SOPs to include additional patient groups (for example, all stem-cell recipients requiring platelet transfusions at our institution).

Conclusion

Initial usage of a SMART-on-FHIR application to optimize platelet transfusion workflow has been consistent with dozens of patients impacted. Work is ongoing to study the clinical impact, for example, transfusions prevented, or prevention of dangerous thrombocytopenia.

Figure 1. Initial usage characteristics of the application. Shown are plots for unique users (clinicians, blood bank staff, etc.) and number of patients loaded into the application (aggregated per month).

References

Assimilating Pollen into Exposomes for Pediatric Asthma Research

Ramkiran Gouripeddi, MBBS, MS¹, Le-Thuy Tran, PhD¹, Tanvi Gangadhar, BTech¹, Randy Madsen, BS¹, Katherine Sward, PhD, RN, FAAN¹, Julio Facelli, PhD, FACMI¹
¹University of Utah, Salt Lake City, UT, USA

Introduction

Pollen is a common trigger for pediatric asthma (PA) exacerbations and other allergic reactions¹. However, there are limited studies that seek to understand the cumulative effects of pollen with chemical species in air quality, and the role of different types of pollen species. In addition, recent reviews of PA have called for incorporation of detailed spatiotemporal dimensions of environmental factors in future studies¹.

Methods

As a part of the Exposure Health Informatics Ecosystem (EHIE)², we have extended the OpenFurther (OF) platform as a comprehensive, standards-based, open-source informatics platform for providing semantically consistent, metadata-driven, event-based integration of exposomic data generated from environmental and physiological sensors, human observations and computational models, with other biomedical data. OF uses an event-driven architecture allowing modeling and storage of all activities related to a study’s operations along with associated metadata in their primitive form on a timeline as events that can be transformed to higher/analytical models based on use-cases.

In order to add pollen data with already existing capabilities of assimilating air quality data with clinical data, we followed these steps. We first surveyed literature for pollen data used in PA research and listed key variables utilized in these studies. We then sampled data available with existing pollen data sources and developed an initial draft of a conceptual model. We reviewed this model with PA researchers and deployed this in a graphical metadata repository.

Results and Discussion

We extended our sensor common metadata specification (SCMS)² to include sensors that measure biological species such as pollen. The extended SCMS captures metadata about (1) instruments used to measure pollen, (2) how these instruments were deployed, and (3) pollen output from these instruments (Figure 1). The output includes concentrations for pollen categories, details about each pollen species, pollen indices, and spatio-temporal coordinates of these measurements. Using this metadata within OF, we generated pollen events representing pollen measurements with spatial and temporal coordinates from two sources: (1) Pollen.com and (2) National Allergy Bureau starting 10/01/2018 for participant locations enrolled in a EHIE pilot study, and are assimilating them with meteorological (supports modeling pollen dispersion), air pollutant and clinical data for PA exposomic studies.

Acknowledgement: Grant Funding: NIH/NIBIB Award: U54EB021973.

References


Figure 1: Conceptual model for pollen metadata and measurements.
Conversational Agents for Chronic Disease Self-Management: A Systematic Review

Ashley C. Griffin, MSPH1-2, Zhaopeng Xing, MPS1-2, Saif Khairat, PhD, MPH1-2, Yue Wang, PhD1-2, Stacy Bailey, PhD, MPH3, Jaime Arguello, PhD, MS1, Arlene E. Chung, MD, MHA, MMCi1-2

1University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; 2Carolina Health Informatics Program, Chapel Hill, NC, USA; 3Northwestern University, Chicago, IL, USA

Introduction

Conversational agents, which are systems that communicate with users in their natural language and are often referred to as chatbots, have the potential to deliver patient-centered interventions that focus on goal setting, feedback, and education. However, little is known how these agents may be utilized to assist patients with chronic disease self-management. Thus, the objective of this study was to review the literature to assess how conversational agents have been used for chronic disease self-management and to identify gaps in the evidence base.

Methods

We conducted a systematic review using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines using five databases (PubMed, Scopus, Embase, Cumulative Index of Nursing and Allied Health, and Association for Computing Machinery Digital Library). A priori search terms included variations of the following: chatbot, virtual coach, dialogue system, and conversational agent. We included full text journal articles or conference proceedings that contained primary research findings for text-based conversational agents focused on self-management for chronic diseases in adults. Studies that did not contain one or more self-management skills1 or chronic diseases as defined by the Centers for Medicare and Medicaid Services were excluded. Two reviewers extracted information from the included studies, analyzed the data, and then grouped studies into themes using an inductive thematic approach. The risk of study bias was rated using the Agency Healthcare Research and Quality internal validity quality evaluation criteria.

Results

We identified 1,606 studies across the databases, and 12 studies met the inclusion criteria. The studies were published between 2012 and 2018, which included five randomized controlled trials, five quasi-experimental, and two non-experimental studies. The duration of studies, excluding laboratory-based studies, ranged from two weeks to four months. The total number of participants across the studies ranged from 10-401, and participants were between 18-92 years old. We rated the risk of bias of one study as good, five studies as fair, and six studies as poor, largely due to the lack of methodological rigor and inconsistent measures. The majority of agents (7) were primarily targeted towards individuals with depression or substance use. Across studies, participants mostly reported a positive attitudes towards agents, and there were improvements on the Patient Health Questionnaire, Generalized Anxiety Disorder Scale, Perceived Stress Scale, Flourishing Scale, and Overall Anxiety Severity and Impairment Scale between the conversational agent intervention and control groups in three randomized controlled trials (p<.05). Major themes that arose from the analysis of user perspectives were related to personalization, engagement, and social support. These themes revealed that 1) participants were interested in tailoring the agents based on their health condition and preferences, 2) graphs, videos, and games facilitated their engagement, and 3) agents have the ability to promote collaboration with support structures.

Conclusion

This systematic review suggests that conversational agents are acceptable to patients and have the potential to support self-management for some chronic diseases. However, this is still an emerging area of research as there is limited evidence for the long-term use of agents at scale. Additional research is needed to assess the influence of more sophisticated dialogue and user models and should use consistent measures for reporting outcomes.

References

Comprehension of Visualizations for Longitudinal Patient-Reported Outcomes

Lisa V. Grossman¹, Meghan Reading Turchioe, PhD, RN², Annie C. Myers², Dawon Baik³, Ruth M. Masterson Creber, PhD, RN²

¹Department of Biomedical Informatics, College of Physicians and Surgeons, Columbia University, New York, NY; ²Department of Healthcare Policy and Research, Division of Health Informatics, Weill Cornell Medicine, New York, NY; ³School of Nursing, Columbia University, New York, NY

Introduction

An increasing number of healthcare organizations capture patient-reported outcomes (PROs) to better monitor patients' health status, as federally mandated by the U.S. financial incentive policy Meaningful Use. Several promising initiatives to design electronic PRO (e-PRO) systems exist, but most do not help patients comprehend their PRO survey results. Helping patients comprehend their results can facilitate self-management, because periodically-repeated PRO surveys can capture changes over time. Typically, researchers use line graphs to help patients comprehend numeric changes over time, including changes in PROs. Unfortunately, previous research indicates that inadequate graph literacy is prevalent, around 40% of the U.S. population. The purpose of this study was to assess patients' comprehension of non-graph visualizations that display longitudinal changes in PROs.

Methods

A purposive sample of 40 hospitalized, English-speaking heart failure patients diverse on age, gender, and race was recruited. We excluded patients with severe cognitive impairment or active psychosis. We assessed demographics, graph literacy, health literacy, and cognitive status. Comprehension (ISO 9186-1) was tested in four conditions, including: (1) text only, (2) text plus visual analogy, (3) text plus number line, and (4) text plus line graph. Each participant viewed every condition, and we systematically randomized the order and information contained in each condition to control for potential learning effects.

Results

Participants had an average age of 61.3 years (SD: 12.5) and were 22% female, 23% Black, 52% White, and 38% Latino. On average, participants scored poorly on graph literacy (mean 1.1 out of 4; SD 1.1), and 37% did not answer any graph literacy items correctly. 55% screened positive for inadequate health literacy, comparable with the U.S. mean. 88% had inadequate cognition (MoCA score < 26). 63% correctly comprehended the text-only condition and 60% comprehended the line graph, compared to 83% for the visual analogy and 70% for the number line (p=0.05). The visual analogy performed significantly better than the text-only (p=0.02) and line graph (p=0.02). Low comprehension was associated with worse cognition (MoCA score 15.7 low vs. 22.3 high, p<0.001), lower education level (p=0.02), and fewer financial resources (p=0.03).

Conclusion

Non-graph visualizations (visual analogies) enhanced understanding of longitudinal PROs compared to text only, but line graphs did not. Our results support using visual analogies rather than text to display longitudinal PROs, but caution against relying on line graphs. Future work will explore comprehension in a larger sample.

Funding

This work was supported by NINR R00NR016275 (PI: Masterson Creber), NLM F31LM054013 (PI: Grossman), and Heart Failure Society of America (PI: Masterson Creber).

References

Prioritization of Clinical Content Informatics Requests Using Multi-Criteria Decision Analysis (MCDA)

Randall W. Grout, MD, MS1,2,3, Dan Seitz, MD3, Justin Morea, DO, MS3
1Children’s Health Services Research, Indiana University, Indianapolis, Indiana; 2Regenstrief Institute, Inc., Indianapolis, Indiana; 3Eskenazi Health, Indianapolis, Indiana

Introduction
Clinical content informatics change requests can overwhelm information technology support and development teams, necessitating a prioritization to optimize their resource utilization. We developed a prioritization algorithm for order set and decision support requests using a consensus rank and multi-criteria decision analysis (MCDA) approach.

Methods
A governance committee (n=17) of informatics, information technology, and clinical stakeholders developed an initial set of prioritization criteria relevant to our organization through a brainstorming and group consensus meeting. They completed an anonymous online survey to rank the items on importance in determining the priority of a request (Figure 1). Consensus ranking was determined via a branch and bound algorithm for a rank correlation coefficient.1 Surrogate weights for criteria were set by the Simple Multi-attribute Rating Technique Extended to Ranking (SMARTER) process and rank order centroid (ROC) formula.2 Scores within each criterion were set as simply linear, equally distributed between 0 and 1. The final prioritization metric was the weighted sum of criterion scores and was implemented in an Excel-based tool for routine use over 1 year.

Results
There was moderate interrater concordance (Kendall’s W = 0.383) in the ordinal rank elicitation. Table 1 contains the final criteria, description, ranks, scores, and weights. Regulatory compliance was an overriding approval mechanism. The committee determined priority scores for 32 significant content and change requests since March 2018, ranging from 0.55 to 0.94 (mean 0.76, SD 0.11).

Table 1. Criteria, scores and weights of MCDA matrix.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Definition</th>
<th>Score options</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>Does this request improve the quality of care?</td>
<td>Yes, Maybe, No</td>
<td>.370</td>
</tr>
<tr>
<td>Workflow</td>
<td>Does this request improve workflow?</td>
<td>Yes, Maybe, No</td>
<td>.228</td>
</tr>
<tr>
<td>Severity</td>
<td>Does this request relate to a severe or high-risk condition or situation</td>
<td>Yes, Maybe, No</td>
<td>.156</td>
</tr>
<tr>
<td>Evidence</td>
<td>Is there strong evidence to support this request?</td>
<td>Yes, Maybe, No</td>
<td>.109</td>
</tr>
<tr>
<td>Frequency</td>
<td>How frequently will a user from the relevant location use this?</td>
<td>Rarely, Monthly, Weekly, Daily</td>
<td>.073</td>
</tr>
<tr>
<td>Quantity</td>
<td>How many patients a month would this be used for?</td>
<td>&lt;10, 10-100, &gt;100</td>
<td>.044</td>
</tr>
<tr>
<td>Finance</td>
<td>Does this request improve the organization’s financial interests</td>
<td>Yes, Maybe, No</td>
<td>.020</td>
</tr>
</tbody>
</table>

Conclusion
We developed and present a MCDA model for prioritization of informatics content and change requests. The weighted scores help allocate and align staff resources with organizational informatics priorities; evaluations are underway. Other institutions may use this decision model or follow the outlined steps to create a model based on local preferences.

References
Extracting the Effectiveness of Physical Activity Interventions from Abstracts

1William Gu, 1Yujing Ke, 1Arthur Zhao, 2Hyeoneui Kim
1Department of Computer Science, 2School of Nursing; Duke University, Durham, NC

Introduction
The health benefits of being physically active are now accepted as a fact. Numerous interventional studies have shown the direct impact of physical activities on disease onset and progress, as well as treatment outcomes.1 However, there has been limited efforts for systematically profiling various physical activity interventions in relation to their clinical benefits by population characteristics. Appropriate intensity and amount of physical activities can vary by patient’s characteristics.2 Linking precise physical activity measures obtained through a personal sensor device with the types, intensity, and amount of physical activities with proven clinical benefits for patients with different socio-demographic background is certainly a crucial milestone towards realizing the precision health initiatives. As an early step toward this goal, this pilot study tested the feasibility of extracting scientific evidence on the effectiveness of various physical activity interventions provided to the patients with Non-Alcoholic Fatty Liver Disease from published abstracts.

Methods
We retrieved 278 lifestyles (including physical activities) related interventional studies on NAFLD patients published in last 10 years from PubMed. Irrelevant studies (e.g., without lifestyle and/or physical activity interventions, study protocol without reported results) were filtered out by reviewing the abstracts. Remaining abstracts were analyzed using Clinical Language Annotation Modeling Processing toolkit (CLAMP).3 To explicitly define and guide the types of information to extract, we first developed a concept model (a.k.a. type system) by analyzing 40 abstracts. In this study, as shown in Figure 1, we focused on extracting relatively simple and straightforward information such as intervention details, effectiveness of the intervention, study sample, and clinical parameters measured to assess the intervention effectiveness. These categories of information and the links that connect these categories are called entities and relations respectively in CLAMP. We trained a text processing algorithm based on this concept model by manually annotating 108 abstracts using CLAMP. The algorithm performance was evaluated with 5-fold cross validation. We defined the type system in such a way that relations are uniquely identified once the two entity types that a relation connects are identified. The key task here was to identify the two entities to pair with a relation. We explored a distance-based approach to relation identification. We developed a rule-based algorithm outside CLAMP that first identified all entity pairs that can be linked with a relation then selected the ones with closest entities.

Results and Discussion
The F-scores of outcome measure, finding, and intervention; which are the core entities of the concept model; were below 0.5 (i.e., 0.44, 0.47, and 0.25 respectively), where precisions were slightly higher than recalls. This result clearly indicates the annotation algorithm needs further improvement. We noted that the inconsistencies in annotation, as small as annotating “ALT (alanine aminotransferase) value” as an outcome measure in one place but only “ALT” in other, hurt the performance significantly. We also encountered common challenges in text mining such as reference resolution (e.g., “it did not improve ALT”, “intervention group showed decrease in body weight”) and omission (e.g., “vitamin A and C”). After correcting the inconsistencies in annotation, the F-score was slightly improved to 0.54. The rule-based relation identification reached the F-score of 0.88. Our follow-up study will focus on improving the performance of the annotation algorithms by training with a larger corpus with better annotation quality. We will consider expanding the corpus to the full text papers especially the method sections to capture the key information with sufficient details. We will also continue refining the type system to incorporate more complex evidence such as comparative findings among interventions, samples, and/or clinical parameters.

References
Researchers’ Perspectives on Symptoms related to Cancer Pain: A Network Analysis of Literatures

Jia-Wen Guo, PhD1, Christina L. Radloff, BSN1, Susan L. Beck, PhD1, Gary W. Donaldson, PhD1, Wendy W. Chapman, PhD1, Katherine A. Sward, PhD1, Lewis J. Frey, PhD2

1University of Utah, Salt Lake City, Utah, USA; 2Medical University of South Carolina, Charleston, South Carolina, USA

Abstract
A network analysis was conducted to explore researchers’ perspectives on symptoms related to cancer pain. Eight symptoms clusters with 24 symptoms were identified. Network structure of symptoms related to cancer pain was similar to that reported by patients, indicating that the available research evidence is applicable to patients.

Introduction
Cancer pain is multifaceted; patients frequently report other symptoms at the same time as they report pain. In order to cope with cancer pain properly, researchers often include other symptoms besides pain in their research. Patients and researchers may have different perceptions of symptoms co-occurring with cancer pain. A network analysis study showed that patients reported cancer pain co-occurring with symptoms of abdominal, urinary, and neurological dysfunction. However, what symptoms researchers included in pain management studies is as yet unknown.

Purpose
To compare researchers’ perspectives on symptoms related to cancer pain with patients’ perspectives.

Methods
Using symptoms related to cancer pain from the literatures as the surrogate for researchers’ perspectives. PubMed, CINAHL, PsycINFO, and the Cochrane Reviews were searched for eligible articles. The search was limited to articles with full-text available in English published between 1990 and 2018. Content analysis was conducted by two nurses using NVivo11 to extract symptoms from the literature. Network analysis was performed in R environment using IsingFit, qgraph, and igraph packages. Data were coded as binary variables, with 0 for absence and 1 for presence of symptoms in the methods or results section of the literature.

Results
Twenty-four symptoms related to cancer pain were extracted and synthesized from content analysis of 144 literatures. To present a network of common symptoms related to cancer pain, 6 symptoms (i.e., body image, hair loss, edema, hot flashes, sexual dysfunction, and weight loss) were excluded in network analysis due to their were coded less than 10 times in the sample literature. The 18 symptoms included in network analysis were: (1) appetite issues, (2) breathing issues, (3) fatigue, (4) sleep issues, (5) anxiety, (6) depression, (7) nausea, (8) vomiting, (9) constipation, (10) diarrhea, (11) cognitive issues, (12) emotional/mental distress, (13) bleeding risk, (14) neurological issues, (15) orthopedic issues, (16) mucositis or dysphagia, (17) skin toxicity, and (18) urinary issues. Eight clusters were suggested (Figure).

Discussion
Clusters found in this study were similar to symptom clusters reported by patients. For example, cognitive and emotional/mental distress in this study were presented as a group of Psychological Symptom Cluster in the previous research. Hormonal Symptom Cluster reported by patients was not presented in this study due to those symptoms were not commonly included in available cancer pain research. Some symptoms were measured collectively in cancer pain research (e.g., emotional distress) while others were measured individually (e.g., anxiety, depression); researchers need to be aware of it when viewing the analysis outcomes.

Conclusion
The similar perspectives between researchers and patients regarding cancer pain symptom clusters indicated that the available research evidence is applicable to patients for cancer pain management.

Reference
Factors Influencing Telehealth Adoption in Critical Access Hospitals: Insights from a Demonstration

Saira N. Haque, PhD, MHSA, FAMIA, Alison Banger, MPH, Sydney DeStefano, BS, Regina Rutledge, PhD, Melissa Romaire, PhD
RTI International, Research Triangle Park, NC

Introduction

Telehealth can expand access to healthcare, particularly in rural areas. However, uptake is low due to several factors including reimbursement and credentialing, among others. The CMS Frontier Community Health Integration Project (FCHIP) Demonstration changes Medicare reimbursement for telehealth, ambulance, and swing bed services at 10 critical access hospitals (CAHs) in frontier communities in Montana, Nevada, and North Dakota over a 3-year period (2016-2019). Eight sites are participating in the telehealth component of the demonstration and receive higher reimbursements (101% of costs) for originating site telehealth visits to increase access to and integration of care. We present facilitators and barriers to telehealth implementation after the first year (2016-2017) of the FCHIP demonstration.

Methods

This work was conducted as part of a larger evaluation of the FCHIP demonstration program. Participating sites were eight CAHs in frontier communities in Montana, Nevada, and North Dakota. We reviewed sites’ applications and progress reports and conducted in-person interviews in May and June 2017 with clinical, administrative, and IT staff. Interview notes were NVivo coded and analyzed using a version of the Consolidated Framework for Implementation Research to identify themes.

Results

Factors influencing telehealth adoption and use include specialty availability, administrative requirements, and change management. The availability of telehealth in relevant specialties can give providers confidence to keep patients in the local community, keeping patients closer to their support networks and reducing travel burden and lost wages. Administrative requirements for credentialing of distant providers in relevant specialties (who provide telehealth services from a second location) also influenced specialty availability. Regional and state-wide shortages of certain specialists, especially in behavioral health, were also cited as challenges. Shortages coupled with low volumes can make it difficult for distant providers to prioritize telehealth visits with CAHs. Administrative challenges included credentialing, identifying workflow processes for telehealth visits and billing and reimbursement. The low volume of telehealth visits can make it difficult for CAH staff to gain familiarity with new systems and protocols. The impact of financial reimbursement was too nascent to assess quantitatively, but was not a primary facilitator because of the low volume of telehealth visits (average of 7 per year in the first year). Change management and implementation support were key factors for telehealth implementation and use. Implementation support for change management included additional funding, training resources, and educational and outreach programs.). CAHs reported that patients were generally receptive to the idea of telehealth visits, especially due to avoided travel time, but knowledge of telehealth services was often low.

Discussion

The findings point to main areas in which those to implement telehealth need support: specialty availability, administrative requirements, and change management. Prior to implementation, identifying specialties for which telehealth services are desired and ensuring that these specialists are available would influence telehealth use. Another area of support is for administrative requirements such as credentialing, reimbursement, and service delivery. The third area of focus relates to change management, including provider acceptance, staffing, workflow and roles and responsibilities.

References

A Scoping Review to Identify Biobank Classification and Permission to Share Metadata

Marcelline R. Harris PhD1, Frank J. Manion PhD1, Madhuri Sankaranarayananpillai PhD2, Hsingyi Song MS2, Cui Tao, PhD2
1University of Michigan, Ann Arbor MI, USA; 2University of Texas Health Science Center at Houston, Houston, TX, USA

Introduction. Biobank classification systems are motivated in part by expressed needs for standardization across biobanks, especially in an era of global collaboration and goals of web-based queries to identify shareable specimens and associated data. Sharing specimens and data requires knowledge about permissions to share that are documented within in consent forms, regulations, and policies. In this project, we explored a) whether the metadata suggested in biobank classification schemes included information about permissions to share, and b) whether the literature on consent to share addressed information about biobanks. This project is part of larger effort to refactor and extend an ontology of informed consent metadata (ICO).

Methods. We applied scoping review methods described by O'Brien (2016). Literature search strategy. A research informationist (MC) searched MEDLINE via PubMed using the Ovid platform, spanning Jan 1 2010 - Feb 23 2018, and using combinations of MeSH terms and keywords: (biobank OR biorepository OR biospecimen resource) AND informed consent; (biobank OR biorepository OR biospecimen resource) AND classification; (biobank OR biorepository OR biospecimen resource) AND informed consent AND classification. Using similar search terms, we later searched Google and Google scholar to identify biobank best practice guidelines, and any articles published after the informationist-aided search. Literature selection process. Abstracts and descriptions of documents were screened for inclusion criteria: English language, full text available, only human specimens and data, and for exclusion criteria: materials were not peer reviewed, did not represent consensus statements, or were not freely available. We then reviewed full text materials. Data collection process. A spreadsheet was used that allowed for online collaboration by the primary reviewers (MRH, FJM, HS) included column headings such as reference, purpose of document, classification categories, terms describing classes and their content, etc. Each document was reviewed by two reviewers, and resolved through discussions with the third reviewer. Synthesis process. A mind mapping tool (Mindjet MindManager 2018) was used to synthesize the data. The classification literature was synthesized first, then best practice guidelines, and finally the consent literature. In this way, biobank classification metadata framed the organizing structure of the mind map.

Results. Our final set of documents for review included 83 articles and 4 best practice guidelines. Not surprisingly, we identified considerable variation across biobank classification schemes. We did not identify meta-level information about permissions for sharing specimens and data in any of the literature on biobank classification schemes, however recommendations about consent information was frequently mentioned as a management and/or quality control process in the best practice documents. We identified two early efforts to ‘ontologize’ permissions for sharing specimens and data, however follow-up with a first author confirmed the dormant nature of that work. We identified several relevant standards-based initiatives, e.g., GA4GH and CDISC.

Discussion
This scoping review identified metadata relevant to both biobank classification and permissions for specimen and data sharing. Interestingly, just as the biobank classification literature did not identify permissions to share as an attribute of the classification scheme, the consent permissions literature did not point to biobank classification systems as context. We are currently in the process of integrating this information into our Informed Consent Ontology and in alignment with other BFO-based ontologies.

Acknowledgements: This work was supported in part by NIH/NHGRI 5U01HG99454 (Tao, PI), and University of Michigan MIDAS data science challenge award (Harris, co-PI).

References

1596
Detection of Opioid Overdose Cases from Text in Electronic Health Records using Machine Learning Methods

Rashed Hasan, PhD1, Mitra Ahadpour, MD1, Henry Francis, MD1, Alfred Sorbello, DO, MPH1
1US Food and Drug Administration, Silver Spring, MD, USA

Introduction:
With the increasing availability of electronic health records (EHR) resources, recent studies at the US Food and Drug Administration are focused on exploring artificial intelligence methods to extract information from text regarding the risk factors for opioid misuse and overdose. To augment opioid overdose case classification from free text clinical narratives, we investigated the application of various machine learning approaches to narratives in discharge summaries using data collected from the Medical Information Mart for Intensive Care III (MIMIC III) database1.

Methods:
We used SQL queries to extract 41,127 text discharge summaries from the MIMIC III database. We used k-nearest neighbors (kNN), decision tree, and artificial neural network (ANN) based machine learning methods to classify the clinical narratives as to their relevance to opioid overdose. For dataset preparation, we used the Natural Language Toolkit (NLTK) set of natural language processing programs to convert sentences to words, remove unnecessary punctuation, dates, and integers, and other related data cleaning tasks. For feature engineering, we used scikit-learn CountVectorizer to prepare the text by converting the patient’s narratives into a bag-of-words model. Each discharge summary was separated into tokens, and the number of times each token occurred in a discharge summary was counted. Then, we used ‘fit transform’ method that helps fit the model to the data then transform the data according to the fitted model. We compiled a set of 129 clinical narratives containing opioid-related terms that were manually assessed as probable or confirmed opioid poisoning cases based on the CDC case definition2 and divided it into a training set of 77 narratives and a test set of 52 narratives using a stratified k fold; a model was trained using k-1 of the folds as training data and the outcome model validated the test data set. Lastly, we applied the kNN minkowski Distance Metric to calculate the distance and find the neighbor overdose class from patient’s discharge summary. The classifier mapped the instances into one of two categories: ‘overdose’ and ‘non-overdose’. In order to measure the quality of the classification, we used a method called the “confusion matrix”.

Results:
As depicted in Table 1, kNN provided the best pattern recognition and classification results of the methods tested.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN</td>
<td>86.11</td>
<td>74.38</td>
<td>79.48</td>
</tr>
<tr>
<td>decision tree</td>
<td>53.29</td>
<td>49.45</td>
<td>51.29</td>
</tr>
<tr>
<td>artificial neural network</td>
<td>49.21</td>
<td>45.33</td>
<td>47.19</td>
</tr>
</tbody>
</table>

Conclusions:
The kNN classifier provided the best performance in classifying the content of text narratives from EHRs into overdose and non-overdose categories. Challenges relate to the multidimensional nature of narrative content. We plan to explore the methodology on larger and more diversified EHR data sets to seek improved accuracy and generalizability.

References

Acknowledgements: This project was supported in part by appointment to the research participation program at CDER administered by the Oak Ridge Institute for Science and Education (ORISE) for the FDA. Funding support received from the FDA/CDER/Office of Translational Sciences. Disclaimer: The views expressed are those of the authors and do not necessarily represent the views of the US FDA or the US Government.
Using Semantic Representations for Urine Drug Tests in the Veterans Health Administration

Corey J. Hayes, PharmD, PhD\textsuperscript{1,2}, Jonathan P. Bona, PhD\textsuperscript{3}, Mathias Brochhausen, PhD\textsuperscript{3}

\textsuperscript{1}Department of Psychiatry, University of Arkansas for Medical Sciences
\textsuperscript{2}VA HSR&D Center for Mental Healthcare and Outcomes Research, Central Arkansas Veterans Healthcare System
\textsuperscript{3}Department of Biomedical Informatics, University of Arkansas for Medical Sciences

Little Rock, Arkansas, United States

Introduction
In 2017 alone, opioid overdoses claimed over 47,600 lives.\textsuperscript{1} Early detection of patients misusing or abusing opioids is essential since receipt of rehabilitation and eventual recovery is unlikely.\textsuperscript{2} One component of early detection of opioid misuse and abuse is through the identification of aberrant urine drug tests (UDTs). Representations of UDTs and their results differ across each of the 170 VA medical centers limiting their usefulness outside of the clinical encounter. With over 125,000 UDTs conducted in one month for cannabis alone, the number of UDTs in VA data is extensive.\textsuperscript{3} This abstract discusses the creation of semantically rich ontological representations of the UDT domain which we are using to continue to provide the foundation for the Opioid Misuse Ontology,\textsuperscript{4} achieve data integration for research, and offer a potential means for early identification of opioid misusers and abusers.

Methods
We provide a basic semantic web framework for representing UDTs using Open Biomedical Ontologies (OBO) Foundry resources. In particular, the Ontology for Biomedical Investigations (OBI) contains classes for a number of similar measurement types, including ‘analyte assay’ (OBI\_0000443) and its subclass ‘measuring glucose concentration in blood serum assay’ (OBI\_0000418), defined as “An assay that determines the concentration of glucose molecules in a blood serum sample.” Its axiomatic definition uses web ontology language (OWL) classes and relations defined in OBI and in other OBO ontologies.

Results
One of the most common UDTs is a 7 panel test measuring the concentrations of marijuana, cocaine, opiates, phencyclidine, amphetamines, benzodiazepines, and barbiturates. To represent tests like the 7 panel UDT, we first define a new subclass of OBI: ‘analyte assay’ with the label ‘measuring drug concentration in urine assay’. To support discovery and use of this term, we also provide the alternative label ‘urine drug test’. We define this as “an analyte assay that determines the concentration of drug molecules in a urine sample.” The axiomatic definition specifies that this assay has as its input (evaluant) a urine specimen, produces measurement data about that specimen, and achieves some OBI ‘analyte measurement objective’. By using explicit machine-interpretable definitions with terms from existing ontologies to represent the involved entities (processes, tissues, molecules, etc), we support full semantic interoperability and automated reasoning about data that would be difficult to achieve with established terminological resources alone. To represent a specific type of drug test, for example the 7 panel UDT, we create a more detailed subclass that specifies the measurement objective in terms of the specific molecules, drugs, or drug classes of interest. When used to represent instance data (corresponding to individual urine drug tests that have occurred and their results), the measurement data item output by this assay is asserted to be about the molecules of interest in the urine specimen.

Conclusion
The variability in the number of ways UDTs and their results are displayed in VA points to the need for standardization of terminology across the VA nationwide. Semantically rich ontological representations of UDTs are a viable option to overcome this variability to improve research and help identify patients misusing and abusing controlled substances particularly by flagging those patients who have aberrant UDTs (e.g. patients positive for oxycodone on UDT but taking tramadol). Future work will include creation of an OWL artifact, publication for public comment, and trialing within a particular VA medical center to better identify patients with aberrant UDT results.

References
Predicting Mental Health Complexity Using EHR and Patient-reported Data

Shan He, PhD¹, Peter J. Haug, MD¹,², Kathryn G. Kuttler, PhD¹,
Brenda Reiss-Brennan, PhD, APRN¹

¹Intermountain Healthcare, Salt Lake City, UT; ²University of Utah, Salt Lake City, UT

Introduction
Integrating care for mental health illnesses into primary care has been proven to improve quality, decrease utilization and lower costs¹. Intermountain Healthcare has developed a team-based care process model for Mental Health Integration (MHI). A MHI packet (9-page questionnaire) is administered to patients who present with possible mental health concerns. The primary care provider (PCP) reviews the packet and assesses a patient’s mental health complexity as “mild,” “moderate,” or “high” based on an internally developed scoring guideline in PDF format. Appropriate mental health treatment plan and resources are subsequently allocated according to the level of complexity. The current process poses a significant challenge to PCPs. The packet generates a large amount of information (68 data items) to review. The scoring guideline is not currently integrated in the Electronic Health Record (EHR) workflow. Previous studies have attempted to computerize the scoring guideline using rule-based models. However, results show only a 53% agreement rate between the computer and physicians². The objective of our study is to apply machine learning (ML) algorithms to the patient-reported packet data combined with features extracted from the EHR to determine the mental health complexity. We have also applied ML techniques to identify patients who are likely to benefit from MHI evaluation with standard screening instruments such as the Patient Health Questionnaires (PHQ-2 & PHQ-9).

Methods
We extracted the MHI packet summary data between 2015 and 2018, crosslinked with 25 variables derived from EHR data (e.g. number of somatic complaints, number of mental health medications, Charlson Comorbidity Index, Area Deprivation Index, etc.) 6 months prior to the packet administration date (7,448 records). We first applied supervised classification modeling to the MHI dataset using physician’s assessment as the reference standard. We then developed a causal probabilistic model³ to identify key components that determine mental health complexity. We also extracted PHQ-9 results documented between 2015 and 2017, crosslinked with variables derived from EHR for the 6 months prior to the PHQ-9 screening date (37,169 records). A dichotomous outcome variable “PHQ9Indicated” was created based on the PHQ-9 total score to indicate whether a PHQ-9 screening is needed (“Negative” vs. “Positive”). Patients without any documented mental health screening (287,321 records) were assigned to “Negative” in addition to patients with a score less than 5. A combination of the WEKA machine learning platform, R, and the Netica Bayesian Network development tools were used to construct appropriate Bayesian Network classifiers and causal inference models.

Results
Predictive modeling using the patient-reported packet data combined with EHR data performed slightly better (66%, AUC=0.74) than results using either MHI data only (64%, AUC=0.72) or EHR data only (59%, AUC=0.64). Our analysis also showed that the individual scoring the patient complexity had the highest correlation with the complexity score, which indicates the limitation of using physician’s assessment as the reference standard for predictive modeling. The causal inference model identified 6 key components of mental health complexity: depression, anxiety/stress disorder, mood/affective disorder, attention-deficit/hyperactivity disorder, family style, and active physical diseases. The causal model was designed to correlate with the assigned complexity and succeeded 74% of the time. The model for predicting whether a patient needs the PHQ-9 screening demonstrated the highest accuracy (94%).

Conclusion
We have embarked on a process that will generate a complexity measure based on a variety of clinical, utilization and socioeconomic status data using both predictive and causal inference modeling. The results provide sufficient encouragement to conduct more extensive ML experiments. The goal is to identify an alternative complexity score that will better predict the therapeutic pathway matching a patient’s needs. We anticipate extending the MHI model and will provide it to all of Intermountain’s caregivers and patients through real-time clinical workflow integration.

References
Parallel Multiple Imputation for Large-Scale Association Studies in Complex Observational Data Settings

Zeling He, MS1,4, David Cheng, PhD2, Chuan Hong, PhD3, Paul Avillach MD, PhD1,3, Isaac Kohane, MD, PhD3, Tianxi Cai, ScD1,3

1Harvard T.H. Chan School of Public Health, Boston, MA; 2VA Boston Healthcare System, Boston, MA; 3Harvard Medical School, Boston, MA; 4Brigham and Women’s Hospital, Boston, MA

Introduction Real-world observational data such as electronic health records (EHRs) are increasingly capturing rich genetic and phenotypic data, making it a valuable resource for large-scale association studies and discovery research. One of the primary challenges in working with real-world data is addressing the extensive and complex missing data that commonly occur due to nonresponse and administrative or data entry error. Imputing the missing data via multiple imputation by chained equations (MICE) is a popular approach to addressing the issue. However, MICE can be highly unstable or computationally slow when applied to a large set of variables (e.g. 1000’s of variables), as required for example in phenome-wide association studies. We propose an alternative procedure that accommodates high-dimensional variables by first screening for predictive variables to use in the imputations and adopting an alternative parallel imputation scheme.

Methods For each variable $X_j$, $j = 1, \ldots, p$, we initially identify variables that are predictive of $X_j$ by testing for significant pairwise association with $X_j$ with correction for multiple testing. We subsequently obtain initial imputations by either a simple mean imputation or a more sophisticated procedure using one-bit matrix completion for binary variables and imputing variables of other types based on the imputed binary variables via regularized regression. Let $\hat{X}_j^{(k-1)}$ be the variables found to be predictive of $X_j$ with missing values imputed by imputations from the $k-1$-th iteration of imputations. The imputations in the $k$-th iteration is obtained by predicted values from a regularized regression of $X_j$ on $\hat{X}_j^{(k-1)}$, which does not depend on other imputations from the $k$-th iteration and can be done in parallel for variables $j = 1, \ldots, p$. This is repeated for $k = 1, \ldots, K$ total iterations until convergence. This entire procedure can be repeated to obtain multiple datasets as in multiple imputation.

Results We evaluated the algorithm in a dataset on autistic children from the Simon’s Simplex Collection (SSC) based on a survey of $n = 2,509$ children that includes $p = 5,459$ features describing demographics, clinical histories, and cognitive and behavioral profiles. The median rate of missingness among all variables was 48.5%. We applied the proposed procedure and calculated out-of-sample AUC and MSE (relative to mean imputation) by withholding 10% of observations for which data was observed, for each feature. The results in Figure 1 show that the proposed parallel procedure achieves comparable performance to that of traditional sequential imputation uniformly in iterations of imputation (cycle over all variables for sequential imputation, 0th round/cycle is initial imputation). Executing the algorithm on 10 parallel threads, the time required to complete 3 rounds of parallel imputations and 3 cycles of sequential imputations were 585 and 5,212 minutes, respectively.

Conclusion Imputations can be computationally difficult to compute in large datasets. We propose an alternative parallel imputation procedure that offers substantial improvements in computation speed without compromising on imputation quality, which can facilitate timely large-scale analyses. We plan to use the approach for large-scale association analyses in EHR data.

References

Application Programming Interfaces in Health Care:
Findings from a Current-State Assessment

Krysta Heaney-Huls, MPH¹, Prashila Dullabh, MD¹, Lauren Hovey, MA¹,
Nithya Rajendran, MHI¹, Adam Wright, PhD², and Dean F. Sittig, PhD³
¹NORC at the University of Chicago, Bethesda, MD; ²Departments of Medicine and
Biomedical Informatics, Brigham and Women's Hospital and Harvard Medical School,
Boston, MA; ³School of Biomedical Informatics, University of Texas Health Science Center
at Houston, TX

Background
Interest in application programming interfaces (APIs) as a means of increasing the ease of health data access for—and exchange among—patients, health care organizations (HCOs), and payers has been growing as federal and non-federal stakeholders look for technical solutions to interoperability challenges.

Research Objectives
We explored three thematic areas to elucidate the current state of API design, implementation, and use: 1) API use cases and standards; 2) challenges and facilitators for read and write capabilities; and 3) outlook for future development of write capabilities.

Methods
Our current-state assessment employed four methods: 1) a literature review of PubMed and the grey literature; 2) expert interviews with 13 API stakeholders (HCOs n=5; electronic health record (EHR) vendors n=2; app developers and third-party data sharing platform providers n=6); 3) a review of three EHR app galleries and the SMART® App Gallery; and 4) a 13 person multi-disciplinary technical expert panel (TEP) to discuss near-term challenges and opportunities to advance API development and use. We used an eight-dimension socio-technical model¹ to organize our findings. The model accounts for technical (e.g., standards) and non-technical (e.g., clinical workflows) aspects involved in the design, development, implementation, use, and evaluation of safe and effective health IT.

Findings
While the use of APIs in health care delivery and research is increasing at an astonishing pace, it is still in the pilot stages, with limited representation in the literature. The use of standards-based APIs to support exchange is an area in which there is much interest, discussion, and development. Experts indicated that where consensus technical standards are used, they are largely Fast Healthcare Interoperability Resources (FHIR®), which is a positive indication of the market’s support for interoperable solutions. A multitude of app marketplaces are becoming available. These marketplaces support a range of use cases, the majority (69% of apps) of which target providers, with far fewer (14%) supporting patient access to data. Results across methods indicate the current focus is on read APIs with very limited use of write APIs. Informants cited reluctance on the part of HCOs and EHR vendors to allow external data to be fully integrated (written) into an EHR. Finally, informants and TEP members agreed that there are inherent costs associated with API development and use that justify some level of fees. We identified five key issues with implications for the continued advancement of API use: 1) a robust normative FHIR® standard; 2) expansion of the Common Clinical Data Set to other clinical data including patient-generated health data and administrative data; 3) enhanced support for write implementation; 4) data provenance rules and guidelines; and 5) costs that are not prohibitive.

Conclusion
This assessment provides a significant window into the current API landscape—including the focus on provider-centric use cases; the rationale for the emphasis on read APIs; and the complexities related to standards. Thus, while APIs are being touted as a solution to the interoperability challenges, they remain an emerging technology that is likely only one piece of a multi-pronged approach to health information exchange, integration, and use.

References
Using Unsupervised Machine Learning Method to Define a New Normal for Body Temperature

Frederik Heath, Myroslava Protsiv, Sarah F. Poole, PhD Julie Parsonnet, MD
Stanford University School of Medicine

Background: Body temperature measurement is a low-cost and non-invasive diagnostic tool that is widely used in clinical practice to diagnose acute pathological states ranging from fever to hypothyroidism. The mean normal body temperature of 98.6 F was established in 1851. The mean body temperature in contemporary, high economy populations, however, is lower than this established standard. We used an unsupervised machine learning algorithm to define normal body temperature in a US population using age, sex, BMI and time of day as parameters.

Methods: The LIMIT Algorithm uses the Central Limit Theorem to mine patient data and establish reference intervals for laboratory tests. We applied LIMIT to outpatient visits (N=615,470) in the Stanford Translational Research Integrated Database Environment (STRIDE). Other variables included patients’ ages, diagnoses, medications, and time of visit. The Hampel method was used to determine if body temperatures (measured orally) were “extreme” (either high or low). Fisher’s exact test was then used to test associations between these extreme values and ICD10 codes or medications. All patients with ICD10 codes or medications that were overrepresented in the extreme temperature tails were excluded from further analysis. In total, 198,924 observations (32.3%) were removed. Diagnoses and medications excluded for high temperatures included acute respiratory infections, acute sinusitis, pharyngitis, bronchitis, pneumonia, viral infections and influenza, as well antibiotics (macrolides, penicillins, quinolones), antivirals and antitussives. Individuals with Type 2 Diabetes mellitus or taking over-the-counter decongestants and antihistimines were excluded for lower temperature values. The remaining population was used to define the normal range of temperature values.

Results: Normal temperature measured at noon assuming age of 30 and BMI of 25 was 97.96 F (95% CI: 97.95- 97.97) for men and 98.14 F (95% CI: 98.13- 98.15) for women. Temperature varied by time of day, age, and BMI (Figure 1).

Conclusion: LIMIT confirmed that normal body temperature is consistently and significantly lower than the canonical value of 98.6°F. Our results are alligned with past findings of the effect of age, sex, BMI and time on temperature, validating LIMIT’s unsupervised a posteriori approach. Accounting for such heterogeneities by using patient specific “precision measures” would allow for better clinical interpretation of body temperature in future.

Using the RE-AIM Framework to Assess the Potential to Use Mobile Diabetes Detective (MoDD) in Federally Qualified Health Canters

Elizabeth M. Heitkemper, PhD, RN1, Arlene Smaldone, PhD, CDE, RN2,3, Suzanne Bakken, PhD FACMI1,2, Andrea Cassells, MPH4, Jonathan N. Tobin, MD, PhD4, Lena Mamykina, PhD1

1Dept. of Biomedical Informatics; 2School of Nursing; 3College of Dental Medicine, Columbia University, New York, NY; 4Clinical Directors Network, Inc., New York, NY

Introduction: Underserved groups (low income, racial/ethnic minorities) with diabetes experience health disparities exemplified by an increased diabetes prevalence, poorer glycemic control, and greater severity of diabetes-related complications [1]. To address growing health disparities, translating innovations developed in research settings into clinical practice has become a national priority [2]. This study used the RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework for clinical informatics interventions to identify the factors necessary for adoption of a novel web-based diabetes technology, Mobile Diabetes Detective (MoDD), in four NYC Federally Qualified Health Centers (FQHCs) that had participated in a randomized controlled trial (RCT) testing its efficacy.

Methods: This mixed methods study was comprised of homogeneous focus groups and individual interviews that included patients randomized to the MoDD clinical trial, health center staff at intervention sites, and study research coordinators. Qualitative data were analyzed using directed content analysis. Quantitative data from the FQHCs’ electronic medical records (EHR) and RCT intervention subject usage of MoDD were analyzed using descriptive statistics.

Results: From November 2015 to May 2017, 15 interviews and 8 focus groups were conducted with three stakeholder groups (patients [N=21], FQHC staff [N=28], research coordinators [N=3]). The majority of patients were female (58%) and African American (52%). Staff were predominately female (79%) and physician (57%). The MoDD usage data (Figure 1) showed that patients engaged with MoDD’s basic features, such as logging into the website and recording blood glucose values; goal setting, a more advanced feature was used less frequently with 41% of patients never setting one. Fifteen qualitative themes were identified corresponding to the dimensions of Efficacy (3 themes), Implementation (7 themes), and Maintenance (5 themes). The themes illustrated how much patients liked MoDD, but how hesitant they were to let providers see their data. FQHC staff stated clear interest in using MoDD, but did not believe that they had the space and staff required for patient training and continued use.

Conclusion: While all stakeholder groups were positive about MoDD, identified patient and health center barriers may limit broad implementation of technologies such as MoDD into practice, regardless of potential efficacy. This study illustrates the current complexity of implementing a web-based technology intervention in underserved groups. For technologies such as MoDD to augment diabetes self-management services in underserved adults, current barriers to implementation must be identified and overcome.

Acknowledgements: This research was supported by NIDDK, R01DK090372; EH was supported by NINR T32NR007969 and Robert Wood Johnson Foundation, grant 73070.

References

Development of an integrated system and central repository for clinical and education outcomes in Latin America

Juan Henao MD\textsuperscript{1,2}, Yuri Quintana PhD\textsuperscript{1,2}, Charles Safran MD\textsuperscript{1,2}

\textsuperscript{1}Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, MA
\textsuperscript{2}Harvard Medical School, Boston, MA

Introduction

Healthcare providers in resource limited settings have difficulties in improving care delivery, in part because of the challenges in systematically collecting outcomes and operational data to improve care interventions\textsuperscript{1,2}. Gains in coverage do not always result in safe and high-quality care due to limitations of training and process improvements.

Methods

Fundación Valle del Lili (FVL) is a 500-bed teaching hospital in Cali, Colombia, that provides training and consultation services to regional clinics with limited resources. To improve care delivery in the region, FVL is developing a community of practice site for regional healthcare providers in Latin America to deliver training, standardize the collection of outcomes, and monitor process improvements. Care delivery processes are being codeveloped and standardized with regional clinics. The goal is to applying learning systems analysis to plan the training, data collection and collect process-oriented data to identify areas for improvement at each center.

Results

A community of practice is being created to provide online access to evidence-based education and clinical best practices. The Alicanto™ is a social learning platform (http://www.alicantocloud.com) is being used that supports online learning communities, providing training, care communication, the collection of clinical, learning and process outcomes, and their review applying learning systems analysis. An online asynchronous discussion forum will be used for communities of practice to share their experiences and discuss challenges in care delivery and data collection with colleagues, and training on tools to collect and analyze that data for quality and process improvement. A group of physicians has been meeting online for the last year to define the educational needs and data collection strategies.

Discussion

Research has the opportunity to link medical education to practice outcomes; however, few educational programs get any substantive feedback on the quality of care delivered\textsuperscript{3}. Fundación Valle del Lili as a teaching hospital has the potential to develop an education data warehouse that combines data of medical school, residency program graduates and their subsequent practice; looking at the differences in value-based care and overall costs through what would constitute the first public database for such comparative analysis in Latin America. This project will also greatly inform new curriculum design efforts of clinical education and training in Latin America. With our comprehensive informatics monitoring framework, we illustrate an innovative continuously learning system that aligns medical education outcomes with clinical process outcomes and builds capacity at the institutional level to generate actionable and comparable knowledge that facilitates analysis, research, and evidence-based decision making in health care, while being scalable to country level in Latin America.

Conclusion

Education, process improvement and quality indicators are interrelated and need a coordinated approach to data collection, analytics, training, and follow-up on process improvements.

References

Data Subset Identification Using CUR Index Selection Schemes

Emily Hendryx, Ph.D.\textsuperscript{1}, Nabil Chaabane, Ph.D.\textsuperscript{2}, Béatrice Rivière, Ph.D.\textsuperscript{2}, Craig Rusin, Ph.D.\textsuperscript{3}

\textsuperscript{1}University of Central Oklahoma, Edmond, Oklahoma; \textsuperscript{2}Rice University, Houston, Texas; \textsuperscript{3}Baylor College of Medicine, Houston, Texas

Introduction

The ability to identify representative subsets of data can be valuable in a number of clinical applications. For example, identified subsets can be used in determining the number of classes present within a data set during a retrospective study, or they can provide a clinician with a summary of the types of physiological waveforms seen within a given time frame. The viability of subset selection using the DEIM-CUR matrix factorization for the primary purpose of class identification and subsequent classification has been demonstrated\textsuperscript{1}. While the DEIM (discrete empirical interpolation method) index selection scheme performs comparably to other methods in analyzing electrocardiogram (ECG) data, the CUR factorization can also be formed using a variety of other methods.

Comparison of Subset Selection Methods

This work compares four different CUR index selection schemes–DEIM, Q-DEIM, leverage scores, and volume maximization–against the more commonly used k-medoids, agglomerative, and Max-Min clustering algorithms. In doing so, we apply all of these unsupervised learning methods to three different freely available data sets: the MIT-BIH Arrhythmia Database of ECG waveforms\textsuperscript{2}, the Grammatical Facial Expressions (GFE) Data Set\textsuperscript{3}, and the MNIST Database of Handwritten Digits\textsuperscript{4}. These labeled data sets present data from a variety of settings, providing examples of physiological waveforms (MIT-BIH), coordinate data extracted from facial images (GFE), and images of handwriting (MNIST); while two of these sets are not directly from a clinical setting, they provide insight regarding expected algorithm performance on similar, yet more clinically-focused, data sets. Each data set is split into training and testing sets for parameter selection for each method of interest. In comparing schemes, we focus on the detection of at least one representative from each known class and the amount of data reduction achieved through subset selection.

Conclusion

While there is not one algorithm that consistently out-performs the others, Max-Min clustering with the dynamic time warping (DTW) similarity measure, Q-DEIM, and DEIM each obtain at least 90\% class detection in all experiments. Although Max-Min clustering achieves the highest class detection rates among these three methods, it does so with the lowest data reduction rate over all experiments at 86.86\% during the GFE analysis and was not allowed to finish the MNIST subset selection task due to excessive runtime. Where the computational complexity of Q-DEIM and DEIM is $O(mnk)$ to select $k$ representatives from $m$ observations of length $n$, the improved class detection results seen in Max-Min clustering with DTW are met with an $O(mn^2k)$ computational cost. With these and other results presented in our work, our experiments show that the CUR index selection schemes perform comparably, if not better than, the evaluated clustering schemes in terms of class detection and data reduction, and often at a reduced computational cost. As the presented methods are extendable to a vast array of biomedical data types, this work suggests that CUR index selection schemes should not be overlooked in data summarization and class identification tasks.

References

Automated Extraction of Computable Clinical Decision Support Rules from an Electronic Health Record (EHR) System

Branden C. Hickey, M.A.S.¹, Marc L. Sainvil M.S.¹, Adam M. Bartscher¹, Robert R. Freimuth, PhD¹, Jane L. Shellum, M.H.A, M.A.S.¹, Davide Sottara, PhD¹
¹Mayo Clinic, Rochester, MN, U.S.A.

Introduction

In 2018, Mayo Clinic replaced many of its clinical applications with a single electronic health record (EHR) system. In the process, several clinical decision support (CDS) rules were built in the new EHR. As manifestations of clinical knowledge, each rule must be validated and endorsed by knowledge oversight groups for patient safety and regulatory purposes. Transparency of rule logic is required to assess correctness, quality, impact and overall outcomes. Unfortunately, the new EHR system does not support standards-based representations suitable for verification and analytics, nor presentation forms accessible to the subject matter experts. We address this gap by means of an automated pipeline that reverse engineers the native implementation of the CDS rules and stores the rule logic and content as discrete data elements. This approach allows the rule to be loaded into a formal content management system and facilitates presentation in human-readable, rather than technical, renderings to subject matter experts (SMEs) and knowledge engineers.

Methods

We analyzed how CDS logic is exposed within the EHR as a large relational model of “configuration parameters”. Snapshots of this database are replicated and cached. We identified a clinically relevant, high impact set consisting of 1400 CDS rules. We implemented a processing pipeline using a C#.NET 4.7 framework: for each rule, information is extracted from the database, mapped to an object model, translated into a HL7 KNART-conformant XML document with CQL/ELM expressions and validated. Using XSLT, the KNARTs are rendered into HTML documents whose style has been informed by a user-centric design. In the process, an ANTLR-based formal grammar is used to analyze the complex logic expressions used in the rules. Coded concepts and value sets are also linked and mapped. The extraction process generates metadata, supporting identification and versioning, so that both the KNART and HTML renditions can be published into a Knowledge Management System (KMS). The stages of the pipeline are exposed through ReSTful APIs, and can be invoked on demand. Rendering of CDS rules will be illustrated in the poster as examples.

Figure 1. Transformation / Publication pipeline

Conclusions and Future Works

At any point in time, the pipeline is able to extract, standardize and publish all the CDS rules and their related value sets. The cache can be updated on demand, but the process takes, on average, less than one second per rule. The combination of the machine (KNART) and human (HTML) readable representations will enable several scenarios. Subject Matter Experts can verify that the actual implementation matches their requirements, and can easily propose updates. Providers impacted by the CDS rules have been able to understand why and how certain recommendations have been delivered. Conversely, reporting and analytics tools can be built to detect changes in the EHR implementation, and identify commonalities. This pipeline adds value to the enterprise by allowing subject matter experts (SMEs) to review CDS rule logic easily which can ultimately impact efficiency and safety.
Using the Lean Startup Method to Efficiently Create an Electronic Consent
Blair S. Hirst, MS, Felix Guerrero, Brendan Mooney, Yuri Turin, Gregory Jordan, Mary Mitchell, John McClean, Daniel Stein, MD, PhD
Memorial Sloan Kettering Cancer Center, New York, NY

Introduction
Procedural consents, one of the most frequently used type of consent forms at Memorial Sloan Kettering (MSK), are routinely conducted on paper. The paper consent process is a manual and time-consuming process and can lead to improperly filed documents, delayed workflows, and ultimately a suboptimal care experience for patients. Conversion from a paper-based to an electronic consenting process presents an opportunity to address issues with the current practice and ultimately to improve clinical care¹. However, converting workflow processes that are longstanding and central to medical center function can be challenging. We are using the Lean Startup method² as an approach to demonstrate how agile practices and data driven validation can produce innovative solutions within our complex, multidisciplinary cancer care setting.

Methods
Eric Ries is the architect behind the Lean Startup method, a Build – Measure – Learn cycle, which advocates for a cyclic and iterative practice in continuous innovation development and we used this method to foster an efficient process and an effective solution for our electronic consent project.

The first step is to generate a hypothesis to solve for a problem through the creation of a Minimal Viable Product (MVP). The hypothesis generated was that an electronic procedural consent will improve clinic and OR space and time utilization, increase patient access and engagement in the consenting process, decrease associated risks with paper consenting, and reduce paper consumption. Our baseline data points included mapping existing consent workflows, identifying usage of consent forms in the Electronic Health Record (EHR), number of consent visits scheduled, and identified usage of consent visits by services over a period of one year. The impact the consenting process has on clinic space and time, as well as a workflow map of the procedural consent clinic visit process was created and will be shown on the poster.

With this information, we began looking for electronic solutions that would address the project needs and be a viable candidate for the Build phase.

Results
Baseline information revealed 26 types of consents used at MSK between August 2017 and August 2018. Of the 26 consent types, only five types accounted for 93.8% of the consenting volume, with the procedural consent accounting for 34.1% of the total consent volume. Consent visits accounted for 2.1% of total clinic visits and a total of five services accounted for 62% of the total procedural consent visits. Over 17000 consent visits where scheduled across 20 services. The average in-room time for consent visits were 69.5 minutes.

The output of our Learn cycle explored and yielded several possibilities, including using existing enterprise systems such as our home-grown patient portal platform or buying off-the-shelf consenting products. In our current iteration, our approach involves integrating a vendor-supplied e-signature platform with our EHR system, using vendor APIs and in-house custom software development. The MVP is currently still being developed and will be used to test our original hypothesis in the clinical setting.

Conclusion
Using the Lean Startup approach, our team successfully generated a working MVP within six months of starting development. The cyclic nature of Lean Startup has allowed for quick technological pivots in the Build phase to ultimately find best solution based on project goals, legal and eSignature requirements. Ultimately, we aim to go beyond consenting in the clinical setting to a mobile and at-home setting based on clinician and patient preferences, possibly integrating with our patient portal.

References
Recurrent Neural Network for Modelling Relevant Sentences for Concept Analysis

Teruyoshi Hishiki, MD, PhD\textsuperscript{1}, Takuro Tamura\textsuperscript{2}, Eiji Sugihara, PhD\textsuperscript{3}, Yoshiko Matsunaga, RN, PhD\textsuperscript{4}, Rie Ichikawa, MD\textsuperscript{5}, Jimpei Misawa, PhD\textsuperscript{5}, Yukihiro Maeda, MA\textsuperscript{5}, Akiko Shibuya, RN, PHN, PhD\textsuperscript{6}, Yoshiaki Kondo, MD, PhD\textsuperscript{5}

Department of \textsuperscript{1}Information Sciences/\textsuperscript{4}Nursing, Toho University, Tokyo, Japan; \textsuperscript{2}Research and Development Center for Precision Medicine/\textsuperscript{3}Headquarters for International Industry-University Collaboration, University of Tsukuba, Tsukuba, Japan; \textsuperscript{5}Department of Health Care Services Management, Nihon University, Tokyo, Japan; \textsuperscript{6}Department of Nursing, Toyama Prefectural University, Toyama, Japan

Introduction

Concept analysis plays an important role in the development of knowledge in nursing science. Rodger’s evolutionary method is one of the well-known methods. Roughly, the method consists of data (i.e., literature) gathering, relevant data selection, and in-depth reading of the data to identify the major theme of the concept and the differences between related concepts. We experienced that these steps are not sequential but reciprocal, because we can refine the original keywords as we repeat searches; theses new keywords would lie in the sentences that are relevant to the working intention (definition) and extension (applicability) of the concept.

Therefore, finding out the relevant sentences to given description of a concept would be a major cognitive activity of concept analysis. However, this would be labor intensive without any computational support, and this computational support was the problem we addressed.

More specifically, the task to be supported was to find sentences relevant to the concept “diagnostic decision making error”. We defined the requirement for sentences to be relevant to “diagnostic decision making error” as follows: a) commentary on the possibility of misunderstanding in the diagnosis of the disease, b) attention to the fact that the disease may exhibit general symptoms or symptoms that seemingly irrelevant to the diseased organs (e.g., skin symptoms due to liver), and c) assessment of a case suggesting the possibility that a misdiagnosis was made, correct diagnosis was delayed, or the disease was not originally anticipated. We manually identified relevant sentences previously. We use recurrent neural network (RNN) to model the word contexts and the inside words of sentences as well as the sentence relevance, and use the model to predict relevant sentences.

Methods and Preliminary results

We made a dataset consisting of 305 abstracts that had been retrieved through PubMed with the keywords “decision making” and “diagnostic error”. We judged that 120 abstracts were relevant to “diagnostic decision making error”, from which we identified one or two consecutive relevant sentences that support this judge. To predict the relevance of a sentence, we modeled the documents following the method of Jagannatha et al (2016).\textsuperscript{1} Word embedding, or vector representation of the words, were done using general-purpose, pre-trained word2vec model.

We divide the dataset into 90% training set and 10% test set and repeat the training and test 100 times. We compare the prediction with our judgment. We also compare frequent words in the predicted sentences with “important words” that we arbitrarily chose from frequent words in the relevant abstracts.

Conclusion

We propose that concept analysis is a reciprocal process rather than sequential, through which keywords taken from relevant documents are enriched. This led us to test recurrent neural network classifier to predict relevance of a sentence to a given concept.

References

**SemMedDB-neo4j: A Graph Database of Biomedical Semantic Relations**

Dimitar Hristovski, PhD\(^1\), Andrej Kastrin, PhD\(^1\), Halil Kilicoglu, PhD\(^2\)

\(^1\)Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia;
\(^2\)National Library of Medicine, Bethesda, MD, USA

Semantic relations from the biomedical literature are increasingly used in knowledge management applications such as literature-based discovery (LBD) as well as in clinical decision-making support. SemMedDB\(^1\) is a repository of semantic predications (subject–predicate–object triples) extracted from the titles and abstracts of all PubMed citations by SemRep\(^2\), a rule-based semantic interpreter. SemMedDB has been so far distributed only as a MySQL relational database. However, for many biomedical applications, the domain knowledge is more naturally represented as a graph of concepts and semantic relationships between them. In this abstract, we describe our recent conversion of SemMedDB to a neo4j graph database (SemMedDB-neo4j).

We generated SemMedDB-neo4j by first exporting the SemMedDB data from the MySQL version (release 31) to several delimited text files. Then, using custom scripts, we calculated additional aggregated data and formatted the data. Finally, we loaded the data into a neo4j graph database.

Some of the data elements provided in the MySQL distribution of SemMedDB are the same as those provided in the neo4j distribution. However, we make some additional data elements available only in the neo4j distribution. These are mostly related to aggregated data and to publication history data. The additional data is especially suitable for discovery and evaluation in LBD.

In general, a Neo4j graph database contains nodes and relations between the nodes. Both the nodes and relations can have properties/attributes. The node types (labels) and relations used in the SemMedDB graph database are as follows: Citation (corresponds to a PubMed bibliographic record), Sentence (corresponds to a sentence), Instance (corresponds to a semantic predication extracted from a particular sentence), Concept (corresponds to UMLS concepts that represent the arguments of semantic predications). The nodes with label “Concept” also have at least one additional label for the corresponding abbreviated UMLS semantic type (e.g. “phsu”, or “dsyn”).

The nodes are linked with the following relation types: IS_IN (a relation between a sentence and a citation), Extracted_From (a relation between a predication instance and a sentence), Inst_Subject (a relation between an instance and the concept which is the subject of that Instance), Inst_Object (a relation between an instance and the concept which is the object of that Instance). Additionally, semantic predicates (e.g., INHIBITS, TREATS) linking arguments are represented as aggregated relations between Concept nodes. For example, there is only one INHIBITS relation between a particular Concept1 and Concept2, but there can be many instances of that relation (Instance nodes).

SemMedDB-neo4j contains ~67M instances of semantic predications aggregated into more than 18M semantic relations. The instances were extracted from ~18M sentences from more than 28M PubMed citations. The graph database is made available at [http://lbd.mf.uni-lj.si/semmeddb-neo4j](http://lbd.mf.uni-lj.si/semmeddb-neo4j). Neo4j provides a powerful query language, Cypher, specialized for graph data, and many well-known graph algorithms, such as shortest paths and community detection.

**Conclusion**

We believe SemMedDB-neo4j will facilitate literature-based discovery methods by simplifying and improving retrieval and aggregation of semantic predications at a large scale.

**References**


1609
Multisource Feedback Driven Intervention Improves Physician Leadership and Teamwork

Jinwei Hu, MD¹,², Robert Lee, MD¹, Sarah Mullin, MS¹, Steven Schwitzberg, MS, MD¹,², Larry Harmon, PhD², Paul Gregory, PhD³, Peter L Elkin, MD¹
¹Department of Biomedical Informatics, JSMBS, University at Buffalo, Buffalo, NY; ²Department of Surgery, JSMBS, University at Buffalo, Buffalo, NY; ³Department of Psychiatry & Behavioral Sciences, University of Miami, Miami, FL

Introduction

Multisource feedback (MSF) provides a method of quantitatively assessing and potentially improving physician teamwork and leadership focused behaviors. First, we sought to determine whether tiered educational interventions can improve measurements of multisource feedback for physician across specialties. Second, whether MSF baseline measurements and improvements after intervention vary by specialty designation.

Study Design

We present survey data obtained over 15 years (2002-2017) using the proprietary PULSE 360 multisource feedback system to measure Leadership Teamwork Index (LTI) scores before and after a tiered educational intervention delivered based on provider scores. PULSE 360 assessments were performed on 1190 physicians from academic (34%) and community hospital (66%) settings in the United States and Canada. 25% of physicians were in primary care, 47% surgical subspecialties and 28% medical subspecialties.

Three tiers of intervention were used depending on baseline LTI score: (1) providing a web based feedback report, (2) debriefing via virtual coaching, and (3) individualized, web based curriculum and directed, continued coaching. Physicians then underwent a post-intervention PULSE 360 assessment 12-24 months after intervention. Statistical analysis was performed using R. We used quasi-likelihood generalized estimating equation models to account for the correlation between subjects across evaluations and assess the LTI score.

Results

Demographics of the 1190 physicians surveyed included 75% male and 24% female respondents. At baseline, there were an average of 2.78 ratings per provider. Surgeons started with the lowest baseline LTI score, and experienced a mean increase of 6.06 LTI points after intervention (p<0.001), whereas specialists had a mean increase of 3.13 LTI points compared to surgeons (p=0.0497).

Conclusions

Baseline multisource feedback scores vary by specialty and improve based on feedback and coaching. In particular, physicians who start with low scores have the greatest potential for improvement.
Cancer Patients’ Emotional Health Discovering: Influences of Health Literacy, Fatalism and Social Support

Pengwei Hu PhD†, Zhaomeng Niu, PhD‡, Shaochun Li
1IBM Research, Beijing, China; 2Rutgers Cancer Institute of New Jersey, New Jersey

Introduction

Cancer patients’ emotional health, which may contribute to their treatment and recovery, is critical to their overall wellbeing. However, cancer patients tend to feel diverse emotional issues such as anxiety, depression, and distress1. Almost 40% of cancer survivors reported having significant distress1. Emotional support from family and friends has been found to increase emotional health of cancer patients2. Cancer fatalism, the belief that death is inevitable when cancer is present, sometimes contributes to negative health outcomes3 and is negatively associated with health knowledge and education4. The current study aimed to investigate the association between health literacy, cancer fatalism, social support, and emotional health of cancer survivors.

Method

Survey data were drawn from National Cancer Institute's Health Information National Trends Survey 5 Cycle 1 (N = 479). HINTS is a nationally-representative survey that aims to collect data on the American public’s knowledge of, attitudes toward, and use of health-related information; and the questionnaires were distributed by mail. Structural equation model (SEM) was used to test the relationship between health literacy, cancer fatalism, social support, and emotional health of cancer survivors and multiple model fit indices were used to evaluate the model fit. Fatalism, social support, and emotional health were used as latent variables and health literacy was used as a manifest variable. HINTS are approved public datasets and are considered as exempt from institutional review board.

Results

Cancer fatalism mediated the effects of health literacy on cancer patients’ emotional health (Figure 1). Health literacy was negatively associated with cancer fatalism (β = -.083, p < .001) while cancer fatalism was negatively associated with emotional health of cancer survivors (β = -.583, p < .01). Social support was positively associated with cancer patients’ emotional health (β = .926, p < .05). The SEM model of the current study had a good fit: χ²(1) = 2.06, p < .001; RMSEA = 0.046; CFI = 0.953.

Conclusion

Health literacy level is negatively associated with cancer fatalism of cancer patients while lower cancer fatalism is associated with increased cancer patients’ emotional health. Therefore, health care providers and family or friend of the cancer survivors need to improve cancer patients’ health literacy and increase social support.

† Authors with equal contributions.

References

Harnessing Health Information Technology in Family Violence in the United States – Systematic review

Vivian C. Hui\textsuperscript{a}, Klem, M. L.\textsuperscript{a}, Rose E. Constantino\textsuperscript{a}, Young Ji Lee\textsuperscript{a,b},
\textsuperscript{a} Department of Health and Community Systems, School of Nursing, University of Pittsburgh, USA \\
\textsuperscript{b} Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, USA

Introduction
Family violence, also dubbed as domestic violence (DV), is a global health vicious cycle that menacing the health and well-being of people around the world. It refers to a pattern of abusive behaviour in any relationship that is used by one partner to gain or maintain control over another intimate partner. Every 9 seconds in the US a woman is assaulted or beaten.\textsuperscript{1} The costs of domestic violence amount to more than $37 billion a year in law enforcement involvement, legal work, medical and mental health treatment, and lost productivity at companies.\textsuperscript{2} Recently, scholarly researchers showed that Health Information Technology (HIT) is an effective way to analyse trends across different populations, assist in developing more effective screening tools and improve the quality of life for victims of family violence.\textsuperscript{3}

Method
1332 Articles are extracted from four databases. 25 research articles met the inclusion criteria. Data abstracted and entered into a table to facilitate the evaluation process in an unbiased, reliable and valid way. Table variables were categorized into three dimensions, namely (1) Classification of intervention (purposes of the study, types of technology, types of intervention and abuse), (2) Descriptive details (sample size, location, and settings), and (3) Outcomes assessment (Measurement, data analysis and interpretation.)

Result
After the systematic review, 4 intervention features and 3 types of intervention outcomes has been found. As for intervention features, 5 mobile applications are included in this review serving as a safety decision, prevention and education aid regarding Intimate partner violence and domestic violence. Another five articles addressed the importance of online learning for healthcare professionals as a training module on how to report DV cases. While 3 articles compared the technology intervention of DV in rural and metropolitan areas, another 16 articles delineated different types of abuse. Elder abuse is the least one supported with technology intervention. Physical and psychological health outcomes are evaluated in 17 articles.

Conclusion
This research elicited data that HIT is effective in alleviating family violence. Future research should devise technology intervention in elder abuse and cyber abuse. More avenues can evaluate the behavioral health outcomes.
Comparison of Data Entry Time for Structured Data Entries and Free Text

Ji Eun Hwang, BS¹, Byung-Ook Seong, MBA², Sang-Oh Lee, MD, PhD²³, Soo-Yong Shin, PhD¹

¹SungKyunKwan University (SAIHST), Seoul, Republic of Korea; ²Asan Medical Center, Seoul, Republic of Korea; ³Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

Introduction
The adoption rate of electronic health records (EHRs) has increased rapidly. However, most current EHRs depend on free text entries, which hinder data usage, especially in research. To overcome this limitation, structured clinical templates or natural clinical language processing are required. Since entering the meaningful data (structured clinical templates) is more important than cleansing the unstructured data (clinical natural language processing), we had interested in structured clinical templates. However, most structured data entries (SDEs) require more data entry time than free text entry¹. We monitored the total data entry time of SDEs and free text in the real clinical practice.

Method
We developed 4 structured pathology reports in the EHR of a Korean tertiary hospital from 2014 to 2015. Within limited time, we used bottom-up approaches to implement structured templates. So, at first, we design the template with physicians, and then SED for template is developed. Although this approach has a demerit of model granularity, it can substantially save the development cost.

We collected the data entry time for four representative pathology SDEs—stomach, lung, colon, and thyroid cancer structured templates—which have been used in clinical practice. We collected log data of SDEs from the deployment of each SDE to 2017 and log data of free text templates from 2011 to the deployment of each SDE. Log data consist of the timestamp from May to July (3 months) of each year (2011~2017) for free text templates or SDEs. Timestamps were collected from 06:00 to 18:00, excluding lunch break (12:00~13:00).

Result
Table 1 shows the comparison results of the average elapsed time between SDEs and free text entries. The interesting characteristics of SDEs with less data entry time are 1) minimizing structured components and 2) input patterns suitable for SDE. For example, the colon cancer SDE has only minimal necessary components based on previous experience, and the thyroid cancer template already has a standardized input pattern, which can be helpful when implementing SDE.

Table 1. Comparison of elapsed time between SDE and free text in the first year of applying SDE.

<table>
<thead>
<tr>
<th></th>
<th>Free Text</th>
<th>SDE</th>
<th>Time comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Elapsed time</td>
<td>Collection year (Number of reports)</td>
<td>Elapsed time</td>
</tr>
<tr>
<td>Stomach cancer</td>
<td>12 min 58 sec</td>
<td>2011~2012 (1,130)</td>
<td>20 min 37 sec</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>14 min 18 sec</td>
<td>2011~2013 (720)</td>
<td>16 min 01 sec</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>16 min 55 sec</td>
<td>2011~2013 (1,002)</td>
<td>16 min 17 sec</td>
</tr>
<tr>
<td>Thyroid cancer</td>
<td>14 min 32 sec</td>
<td>2011~2014 (1,574)</td>
<td>12 min 50 sec</td>
</tr>
</tbody>
</table>

Conclusion
When designing SDEs, we should focus on the reduction of data entry time for successful deployment. As in the case of colon and thyroid cancer, well optimized and designed SDEs will reduce the elapsed time of data entry.

References

1613
An Informatics Framework for Knowledge Representation and Reconciliation of Disease Pathways

Yulia Innokenteva¹², Olha Kholod¹², Fei He³, Duolin Wang¹, Richard Hammer²¹, Dong Xu³¹ and Dmitriy Shin²¹*

¹MU Informatics Institute, ²Department of Pathology and Anatomical Sciences, ³Department of Electrical Engineering and Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, ⁴School of Information Science and Technology, Northeast Normal University, Changchun, 130117, China

Introduction

The practice of precision medicine requires fine-grained knowledge of biological processes to explain patient-specific clinical phenotypes. However, the vast majority of current knowledge bases (KB) contain canonical disease pathways and rarely include uncommon disease mechanisms and drug actions. Furthermore, state-of-the-art knowledge representation (KR) methods still lack a capability to enable uncovering of implicit relationships without applying more complex reasoning techniques. To address these challenges, we developed a novel informatics framework for KR and reconciliation of disease pathways for precision medicine analytics. The framework introduces a multi-perspective consensus method for processing concordant, contradictory, and complimentary disease mechanisms.

Method

To represent a disease pathway (DP) on the biological level we introduced a notion of a Disease Pathway Profile (DPP). There can be several profiles for the same disease pathway. A profile can be viewed as a specific perspective of a disease pathway. For instance, a newly reported disease pathway that describes a downstream effect of a genomic alteration such as a Single Nucleotide Polymorphism (SNP) in EGFR gene on its 20th exon known as T790M in Non-Small Cell Lung Cancer (NSCLC) can have several profiles. One profile can describe genomic alterations involved in the disease pathway and can be designated as DPP-Genomic_Alteration, another profile can describe regulation by the disease pathways of a specific cancer hallmark (cell survival, proliferation and differentiation or/apoptotic mechanism) and can be designated as DPP-Cancer_Hallmark. On the technical level, a disease pathway is represented as a network, where nodes correspond to molecular entities such as genes, proteins, chemical elements as well as molecular processes such as metabolic reaction, and links constitute interactions between these entities and/or processes. Nodes and links are linked to ontological terms. An elementary network unit is an interaction between a pair of molecular entities and/or processes (e.g. “kinase A phosphorylates protein B” or “metabolite C participates in metabolic reaction D”). When disease pathways from different articles can be meaningfully linked, an integrated disease pathway profile is formed. The main logic of the proposed multi-perspective consensus model is as follows. For each newly extracted disease pathway, the KB is queried to determine the status of the pathway. New disease pathway: If there are no profiles in the KB that exactly match the extracted disease pathway from the same perspective, then a new profile is created and the process is finished. Concordant disease pathway (aggregation): If the extracted disease pathway exactly matches a profile in the KB, it is aggregated with that profile and the profile’s evidence score is increased. Contradictory disease pathway: If the disease pathway partially matches a profile in a KB, a new profile is created, which constitutes an alternative annotation to the existing profile(s). Complementary disease pathway (integration): If there is an existing complementary profile in the KB, it is augmented by a newly created DPP and a new integrated disease pathway profile is formed and entered into the KB. We used an RDF store as the KB platform to store extracted disease pathways and provided retrieval capabilities using SPARQL language.

Results and Conclusion

We have tested our framework on a set of disease pathways reconstructed from 195 PubMed articles. Our method enabled formalization and inclusion into a KB a large spectrum of complex and heterogeneous biological mechanisms. Moreover, the multi-perspective model of the framework facilitated uncovering of a common mechanism of V600E mutation in BRAF gene in Melanoma and Hairy Cell Leukemia disease pathways in an explicit way. That link was previously reported in the literature and consequently validated. We, therefore, conclude that the proposed framework can improve health outcomes in individualized patient cases by efficiently bringing the latest biomedical discoveries into a precision medicine setting.
Mapping Medications to Standards with Machine Learning and Noisy Labels

Alvin D. Jeffery, PhD, RN, 1,2 Sharidan Parr, MD, MS, 1,2 Michael E. Matheny, MD, MS, MPH 1,2
1. TN Valley Healthcare System, U.S. Department of Veterans Affairs, Nashville, TN, USA
2. Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

Introduction

Medication information is commonly used in large-scale, observational studies. Attempts to map medications to standard terminologies have explored information extraction from clinical notes, order sets, problem lists, and other clinical records. However, mapping medications within large, multi-site databases is not routinely performed, and if it is completed, it might not be correct. Recent, novel algorithms to perform such medication mapping tend to be conducted in condition-specific settings (e.g., *C. difficile* infections) rather than in a generalized manner. Within the Department of Veterans Affairs (VA) national data, as many as 15% of prescription orders are not mapped to a standard terminology, and up to 60% of prescription orders cannot be directly mapped to a standard RxNorm concept unique identifier (CUI). Noisy labeling (i.e., using imperfect but available labels) combined with machine learning could provide a process for efficiently and accurately mapping medications to a standard terminology. Therefore, we sought to evaluate the performance of an algorithm that uses noisy labels to map medication data to RxNorm CUIs.

Methods

We developed a multi-class random forest classifier to predict generic RxNorm CUIs within observational data from the VA. We randomly selected 5 medical centers for preliminary testing and applied the classifier to initial prescription data for all patients with at least 2 encounters. We aggregated individual prescriptions by standard identifiers, drug names, strength, unit, and form, and calculated summary statistics for medication quantity, days-supply, refills, and price. We then converted the VA-specific identifier (VUID), the National Drug Classification (NDC) identifier, and the ingredient/strength/unit/form (ISUF) combination to RxNorm CUIs via RxNav. The resulting CUI served as the outcome in each model with the number of unique labels ranging from 5,739 (NDC-based) to 17,293 (ISUF-based).

We removed aggregated rows representing < 5 prescriptions from 2005 to 2018. We included all 43 numeric features (see Table) as predictors in each model. Each model also retained categorical predictors for the original VUID, NDC, or ISUF values (see Table). We used median imputation for numerical features and imputed the word “MISSING” for categorical features. Our hold-out test set comprised 100 manually- adjudicated cases reviewed by a nurse practitioner.

Results

All models yielded Micro F1 scores ranging 0.77-0.89 (see Table), indicating moderate to strong performance. The highest-performing models used the VUID as a predictor and either the VUID-based generic CUI or NDC-based generic CUI as the outcome. Accuracies ranged 0.71-0.89. In most models, the strongest predictor was the standard term (e.g., VUID to predict VUID-based CUI). The number of individual prescriptions an aggregated row represented (i.e., *counts*) and minimum price were frequently found in the top 5 strongest features.

Discussion and Conclusion

We have demonstrated the feasibility of using machine learning and noisy labels to map prescription-based medication information found in a large databases to a standard terminology. We achieved moderate-to-strong performance on a small set of manually-adjudicated cases. We plan to expand the manually-adjudicated test set, train the model on all 130 VA medical centers, and assess model performance on rows with unlabeled data using the manually-annotated test set. These final results will be presented at the AMIA conference.

References

Utility of the Business Process Model and Notation Standard to Represent Business Processes in the Arden Syntax

Robert A. Jenders, MD, MS, FACP, FACMI
Charles Drew University & University of California, Los Angeles, CA

Introduction

Arden Syntax is a formalism supervised by Health Level Seven International (HL7) for representation of procedural medical knowledge with the goal of sharing units of knowledge known as MLMs. An MLM, which may contain business processes, consists of 3 required categories: Maintenance (software attributes of the MLM such as author); Library (with Purpose and Explanation slots for unstructured descriptions of what the MLM does); and Knowledge (with Data, Logic and Action slots for explicit, structured encoding of executable knowledge). While Arden Syntax has a rich set of features for representation of clinical knowledge, it lacks explicit constructs for business process modeling. Sites using Arden have constructed workarounds to do this, such as EVOKE statements invoking EVENT mappings or WRITE statements to represent a business event, but these can end up incomplete, omitting details that are left implicit. Business Process Model and Notation 2.0 (BPMN) has been suggested representation of workflows, reflecting joint work with the Object Management Group (OMG) to use combinations of languages to represent clinical processes. BPMN represents entities such as events (things that happen, such as escalation and conditional), tasks (which are atomic activities that are part of processes) and sequence flows among them. Prior work has demonstrated the significant prevalence of business processes in MLMs, although commonly in narrative format or implied. In order to explore the utility of including the BPMN standard as part of the Arden Syntax, the present work assesses the representational adequacy of BPMN for business processes in the Arden Syntax.

Methods

A convenience sample of MLMs pooled from 5 sources—3 university hospitals and 2 vendors—was examined. Business processes, either structured in the Knowledge category or unstructured in the Maintenance category of each MLM, that align with constructs in BPMN 2.0 (such as events, activities and workflows) were identified.

Results

A total of 325 MLMs were examined. MLMs concerned mainly with lab tests were the most common (137/325 = 42%), followed by clinical assessment or classification (67/325 = 21%) and medication (41/325 = 13%). The remainder addressed administrative and miscellaneous topics. These included 430 trigger mappings that could be modeled as an Event such as Escalation or a Conditional Event, but often with incompletely defined workflow (e.g., responsible clinician, escalation pathways, etc). The most common entities of the Event type were hospital admission and a laboratory test result, while ordering of a medication or a laboratory test was the most common Task. In addition, 1027 output definition statements were identified, and these were representable by the Choreography Task element. For example, an alert issued via an Arden WRITE statement in the Action slot can be expanded into a Choreography Task, with explicit representation of the recipient role as Participant and the recommended action via a Message and messageFlowRef. Also, processes referenced in the Purpose and Explanation slots in the Library section of nearly all MLMs could be represented by the Choreography Task and Sequence Flow elements of BPMN. There were no business processes that could not be represented in BPMN.

Conclusions

Business processes are a common feature of knowledge represented in Arden Syntax MLMs. These processes are represented in part through explicit data mappings of events, activities and message output destinations. However, more nuanced processes are present only in narrative format because of the absence of specific formal constructs that would support their representation. BPMN is sufficiently robust to represent these processes in a computable format in the Arden Syntax. While it would be the subject of future work in HL7, as BPMN lacks an expression language of its own, it could be integrated in Arden through new native Arden operators or, as a version of the standard already exists in this format, via amendments to Arden’s XML schema.

References

Acceptance and Use of Home-Based Electronic Symptom Self-Reporting Systems in Cancer Patients: A Literature Review

Yun Jiang, PhD, MS1, Huiting Zhang, MSN1, Marcelline R. Harris, PhD1
1University of Michigan, Ann Arbor, MI, USA

Introduction
Electronic systems developed to support cancer patients’ self-reporting of symptoms from home have shown promising potential to support cancer symptom self-management. However, patients and families’ adoption of these systems, especially their long-term use of the systems, has not been fully understood. This purpose of this review was to gain an overview of cancer patients’ acceptance and use of electronic systems for symptom self-reporting from home, and to identify factors associated with patients’ acceptance and use of the systems.

Methods
The electronic database PubMed, Scopus, PsycINFO, CINAHL, and Cochrane library, from January 2003 to December 2018 were searched using the keywords: “self-report”, “patient reported”, “symptom reporting”, “adverse effect”, “telemedicine”, “mobile health”, “ehealth”, “technology acceptance”, “technology adoption”, and “cancer”. English articles were selected if the studies assessed adult cancer patients’ acceptance and use of electronic systems for symptoms self-reporting from home. Electronic systems only used in clinics or in clinical trials were excluded.

Results
A total of 569 articles were retrieved and 25 articles met the study criteria; most were short-term feasibility tests with small sample size (range 5–286). Among 17 electronic systems studied, the majority were internet-based (e.g. STAR, WebChoice, InCASA), and 4 mobile phone-based (e.g. ASyMS, Interaktor) and 4 Interactive Voice Response (IVR) telephone systems. Studies included patients with multiple types of cancer that self-reported symptoms related to chemotherapy, radiotherapy, or surgery. This review revealed inconsistent definition or measurement of patient acceptance and use of electronic symptom self-reporting system across the studies. Acceptance was assessed by patient participation rate, positive expectation, general satisfaction, perceived usefulness or easy-to-use, and intention to use. Actual use of the system was assessed by the percentage of participants that used (logged on) the system as intended, completion rates (of the study), full or minimal compliance (patient-days with all data or at least one parameter available), daily or weekly call adherence, etc. Inconsistent assessment makes the synthesis of findings difficult. In general, patients reported high satisfaction, positive perception of ease of use, and minimal technical issues such as internet connectivity. Four studies followed up with patients for 12 months or longer, indicating the greatest use in the initial 12 weeks; however, the compliance declined over time; and with a mean follow-up 34 weeks, 83% monthly compliance (self-report at least once per month) and 62% weekly compliance were reported; The studies with 1-year follow-up showed 77% patients logged on at least once and 63.6% logged on more than once. The greater use was associated with older age, higher education, high level of computer experience, and not having comorbidities. Common reasons for patients’ non-use of the system included perceived no benefits, feeling too sick, feeling too distressed, unwilling to use a computer, limited health literacy, high frequency of reporting, too many reporting questions, and no symptom to report. Facilitators of patients’ acceptance and use included: timely feedback from clinicians, peer connection or interaction, alert or reminder of use, user friendly interface, easy to access or use, informative content, linkage to self-management resources, and adequate health literacy.

Conclusion
Current studies lack the standardized assessment of patients’ acceptance and use of electronic symptom self-reporting system. More research is needed to increase our understanding of patients’ long-term adoption. A variety of barriers to patient use of self-reporting systems indicated the level of tailoring needs need to be improved, in order to develop personalized technology-based interventions to support cancer patient symptom self-reporting.

Acknowledgements: This study was supported in part by the National Institute of Nursing Research (P20NR015331).

References
Review of Mobile Applications for the Detection and Management of Atrial Fibrillation

Victoria Jimenez, Samuel Isaac, Munther Alshalabi, Ruth Masterson Creber, PhD, MSc, RN, Meghan Reading Turcione, PhD, MPH, RN

Weill Cornell Medical College, New York, NY

Introduction: Atrial Fibrillation (AF) is a cardiac arrhythmia affecting an estimated 6 million people. Advances in smartphone technology and mobile health (mHealth) have made it possible for patients to monitor and manage their cardiac conditions at home. However, there is little guidance for patients and clinicians on which available mHealth applications (apps) are appropriate for managing AF.

Objective: This review of commercially available applications aimed to identify and assess the functionalities mHealth apps for monitoring and managing AF leveraging native smartphone capabilities.

Methods: We used an established methodology for this review. We identified relevant apps from three marketplaces: AppStore, Google Play Store, and Amazon Appstore with standard search terms (e.g. “Atrial fibrillation”, “ECG”, “Heart rhythm”, “Cardiac arrhythmia”). Apps underwent four rounds of screening, designed to identify AF specific patient-targeted apps (Figure 1). We evaluated the final list of included apps using the Mobile Application Rating Scale (MARS), IMS Institute for Healthcare Informatics functionality scores, and evidence-based guidelines. The MARS assesses the broad quality and performance of apps; while IMS functionality score evaluates the app against 11 functionalities. Evidence-based guidelines for general AF management were used in the absence of patient-specific self-management guidelines.

Results: We screened an initial 1,473 potential apps, and 13 met the final inclusion criteria. Apps were largely excluded based on content aimed at a clinical, non-patient audience; or content solely for entertainment purposes. While a portion of apps provided pulse data, they failed to measure cardiac rhythm (n=70); and they did not provide any further education regarding AF. The average MARS score of the included apps was 3.0 out of 5. Nine of 13 apps had scores of 3.0 or greater, which is the minimum acceptability score. There was high interrater reliability among the four reviewers (IRR=0.75-0.83). Of the apps evaluated, Afib Companion had the highest overall MARS score (4.1), followed by Qardio and Photo Afib Detector (both scored 3.6). The most common functionalities were record, display, and collect data; allowing patients to track cardiac rhythm and medication adherence data, and view trends in self-reporting (Figure 2). None of the apps adhered to all the evidence-based guidelines for AF management.

Conclusion: Despite the volume of relevant search results, few apps are actually leveraging native smartphone capabilities to provide patients-facing tools and education regarding AF. The apps identified had adequate quality scores but limited functionality, and did not adhere to evidence-based guidelines. Future app development and improvement of existing apps should aim to adhere with guidelines for AF management and offer greater functionality to better meet the self-management needs of this patient population.

References
LeukApp: Development of a Mobile Application for Clinical Decision Support in Acute Leukemia

Andrew Johnson¹, Stephanie Furniss, PhD¹, Matthew Burton, MD¹, Priya Radhakrishnan, MD², Jeffrey Schriber, MD², Anita Murcko, MD, FACP¹

¹Biomedical Informatics, College of Health Solutions, Arizona State University, Scottsdale, AZ; ²HonorHealth Virginia G. Piper Cancer Care Network - Cancer Transplant Institute, Scottsdale, AZ

Problem Description: Optimal and timely treatment of acute leukemia requires a complex evaluation of a patient’s prognosis and suitability for hematopoietic stem cell transplant (HSCT). HSCT offers the highest likelihood of long-term disease-free survival compared to standard chemotherapy and radiation treatment. However, HSCT is a more aggressive treatment pathway that also confers risks of treatment-related mortality and adverse events, especially in older patients, patients with multiple comorbidities, and patients with decreased functional status¹. Additionally, acute leukemias are highly heterogeneous diseases, which present varying risk categories based on clinical factors, French-American-British (FAB) classification, molecular and cytogenetic factors, as well as prior relapse or treatment². Given the difficulty and importance of timely diagnosis, complete testing and correct treatment, HonorHealth is developing a mobile application to enable clinicians to utilize clinical data to understand a patient’s prognosis and facilitate prompt referral for HSCT if indicated. This is part of a larger initiative to develop a specialized leukemia learning network, an educational program to support the inter-professional team at community based hospitals and critical access hospitals to improve care of patients with acute leukemia.

Methods: To develop the mobile app, HonorHealth partnered with clinical informaticists from Arizona State University. Eleven semi-structured interviews were conducted with emergency physicians, hematology/oncology specialists, hematopathologists, and primary care providers. These interviews were used to create workflow models outlining pathways for patient and information flow through initial work-up and leukemia diagnosis, specialized evaluation, treatment, and follow-up. These findings were used to generate use cases and identify target users for the proposed application.

Discussion: Based on our evaluation of interviews with physician experts, we identified two key use cases for the proposed app: (1) evaluation of a patient’s prognosis into key risk groups - favorable, unfavorable, and intermediate; and (2) determining a patient’s potential eligibility for HSCT based on demographic, clinical, and cytogenetic factors. We determined that oncologists and hematopathologists would be the target users, based on the physician’s level of involvement with reaching treatment decisions after the initial diagnostic work-up. The application would incorporate a knowledge base and implement a rule-based approach for assigning treatment prognosis and determining HSCT suitability using decision rules from the above factors. We plan to gather feedback on a prototype of UI mock-ups and screen maps using a focus group of nurses and physicians from HonorHealth and other institutions.

References
The impact of an automated digital navigation program for colonoscopy on bowel preparation quality and patient satisfaction: a comparative study

Sarthak. Kakkar, MS¹, Rajanshu. Verma, MD², Hallie. Jara, Degrees², Shashank. Garg, MS¹, Andy. Pfau, JD³, Cera. Jaffe, BS³, Shrawan. Patel, MBBS, BSc³, Virender. Sharma, MBBS², Ashish. Atreja, MBBS, MPH¹

¹ Icahn School of Medicine at Mount Sinai, New York, NY, United States; ² Arizona Center for Digestive Health, Phoenix, AZ, United States; ³ Rx.Health, Secaucus, NJ, United States

Abstract

About 1 in 5 patients suffer from inadequate or poor bowel prep in advance of a colonoscopy.¹ Prior studies have shown that this leads to a decrease in adenoma detection rate, increased time spent in completing colonoscopy procedure, as well as an overall increase in cost of care from 13-20%.² There has always been a patient need to improve this aspect of care, however, with the shift towards value-based care in the American healthcare system, there is now a business case for urgent improvement. Digital care pathways linked through SMART on FHIR have potential to guide patients through complex periprocedural guidance in an intuitive manner.

OBJECTIVES To integrate and utilize a digital navigation program to automate and improve preprocedure navigation for colonoscopy and ascertain its impact on bowel preparation quality. METHODS A quasi experimental design was used to compare two cohorts of patients (usual care versus digital navigation) scheduled for colonoscopy at the Arizona Center for Digestive Health (AZCDH). The digital navigation cohort received usual care in addition to clinical rule driven time based messages leveraging specific content. These messages utilized multimedia education modules and were built on, and delivered through, RxHealth’s digital medicine platform. Patients received notification messages and education content from the day of appointment scheduling right through to one day post procedure completion. We then analyzed the characteristics of bowel preparation in the two cohorts, and patient satisfaction with the digital navigation intervention. RESULTS Of the 217 patients prescribed the digital navigation program, 20 (9.21%) patients had text messages undelivered (due to wrong number or landline), and 93 completed the procedure and had bowel preparation results documented in AZCDH's electronic endoscopic record system. Patients who received the digital navigation program (n=93) were significantly more likely to have good or excellent bowel preparation quality than patients in usual care (n=960, 87.10% vs. 64.79%; p<0.05). As shown in Figure 1, the digital navigation program cut down the rate of aborted procedures by more than half (1.07% vs 2.2%). Furthermore, 27 patients responded to the follow-up electronic satisfaction survey and 92.59% of them “strongly agreed” and “agreed” that the digital navigation program was helpful in preparing them for a colonoscopy.

CONCLUSIONS Our single site study showed that digital navigation for colonoscopy is well adopted by patients and can lead to better bowel preparation rate and significant reduction in rate of aborted procedures, while achieving high patient satisfaction. Future research should evaluate this in multisite trials and identify strategies to engage patients who are not digitally engaged in their care or lack the skills or resources to do so

References

Data-driven Study of Pain, its Predictors, Co-morbidities, and Characteristics using Nurse-generated Big Data

Youjeong Kang¹, PhD, MPH, CCRN;; Durga Prasad Sanugula², MS; Kriti Bagdi³, MS; Robin R. Austin⁴, PhD, DNP, DC, RN-BC; Karen A. Monsen²-⁴, PhD, RN, FAMIA, FAAN

¹ College of Nursing, University of Utah, Salt Lake City, UT
²Computer Science and Engineering, ³Institute for Health Informatics, ⁴School of Nursing, University of Minnesota, Minneapolis, MN

Introduction. Previous studies in health care industry have made use of big data across different clinical settings, disciplines and data sources. For example, nurse researchers have used the Omaha System, a multidisciplinary ontology and terminology for big data analysis to identify critical data elements in nursing documentation (1). The Omaha System includes three components designed for assessment (Problem Classification Scheme), care plan/services (Intervention Scheme), and evaluation (Problem Rating Scale for Outcomes) (1,2). Under the assessment component, there is a taxonomy of 42 health concepts called the problems including signs and symptoms used in clinical assessments (1,2). Although there is a long history of research using the Omaha System, the pain component has not been fully utilized. In general, there is a lack of pain research using big data approach (3). Thus, the aims of this study were to 1) using machine learning techniques, identify primary and secondary predictors of pain, 2) examine the likelihood of having pain across the population based on most common predictors identified in Aim 1.

Methods. This study used existing de-identified nurse-generated Omaha System data within the secure data shelter of the University of Minnesota Academic Health Center. The data include Problems, Interventions as well as Knowledge, Behavior, and Status (KBS) scores. Pain variables included 6 binary signs/symptoms, interventions addressing the Pain problem, and Pain KBS on admission and discharge. The Problem Rating Scale for Outcomes consists of three Likert-type ordinal scales for rating problem-specific KBS from 1 (most negative) to 5 (most positive). For Aim 1, we trialed Machine Learning algorithms like Logistic Regression with L1 regularization, Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), Decision Trees (DT), Gradient Boosting (GB) and Random Forests (RF). Tree based models performed better as the data with 325,369 samples is very high-dimensional and the target variable was computed by having some conditions on independent variables. We used 80-20 train test split and 5 fold cross validation to determine the best model. Based on the following validation accuracies of various ML models (Decision Tree: 99.31%, Gradient Boosting: 98.31% and Random Forest: 95.01%) we chose Decision tree for extracting primary and secondary predictors of pain and derive optimal models and interpret findings specific to pain. These models were built using scikit-learn ML library. We used descriptive and inferential statistics to examine pain for cases with pain and other co-morbid problems.

Results. The sample (N=3,934) include 58.2% of female and 57.9% of white with a mean age of 74 years (SD=17.6). The strongest primary predictors of pain (normalized feature importance score=0.1) were the sign/symptom expresses discomfort/pain and the baseline Pain Status score. The strongest secondary predictors of pain were decreased muscle strength and gait/ambulation disturbance, both signs/symptoms of the Neuro-musculo-skeletal function problem, overall Status score at discharge, and the total number of signs/symptoms. For the entire dataset (N=325,369) cases with the Neuro-musculo-skeletal problem were 454 times more likely to have Pain vs. cases without the Neuro-musculo-skeletal problem; while patients with the Circulation problem were 204 times more likely to have Pain vs. cases without the Circulation problem.

Discussion. This data-driven, comparative study addressed the problem of understanding pain from a large nurse-generated dataset perspective overall, and for sub-populations within the big dataset. Results revealed new patterns in pain predictors. Future study needs to be replicated in other datasets with smaller dimensions to validate findings because it is easy to lose the important vector if the data are very high-dimensional. Further research is underway to examine pain, multi-morbidity, intervention effectiveness, and differential outcomes.

Conclusion. Little research has been conducted to explore pain using standardized nursing terminology data. This study demonstrated the potential for knowledge discovery related to pain within large nurse-generated datasets.

Life after cancer: After-Treatment concerns and peer-support among breast cancer survivors, portrayed on Twitter

Akshat Kapoor, PhD, MS\(^1\), Elizabeth Baker, MS\(^1\)
\(^1\)East Carolina University, Greenville, NC, USA

Introduction

The rigorous cancer treatments, including surgery, chemotherapy, and radiation can often have a debilitating impact on one’s physical, psychological, and social well-being. There is ample evidence on the benefits of peer-support interventions for cancer survivors, however, knowledge about the role and characteristics of peer support in an unmoderated online environment is still limited\(^1\). Analyzing tweets of breast cancer survivors can thus provide a meaningful insight into how cancer treatment-related symptoms impact the daily lives of survivors, and the various coping mechanisms they adopt to maintain their quality of life.

Objective

The objective of this research is to answer the following research questions: RQ1: What are the primary post-treatment concerns among breast cancer survivors as shared on Twitter?; and RQ2: What are the characteristics of peer-support and associated information-sharing among breast cancer survivors on Twitter?

Method

We utilize a random sample of 100 Twitter users comprising of breast cancer survivors. Survivors were identified as having tweeted about receiving breast cancer treatment at least once on their Twitter timeline. Only individual accounts were included, and organizations were excluded. We perform content analysis on the survivors’ original tweets (retweets were not included in the analysis) to identify post-treatment concerns and forms of support, as depicted in tweets. Concerns were further categorized as physical, psychological, or social. Similarly, forms of support in tweets were labeled as emotional, instrumental, informational, or appraisal support. Thematic analysis was performed to identify recurring themes. Tweets were aggregated using the Twitter4J Java library for Twitter API, and qualitative analysis was performed using NVivo.

Results

Approximately half a million tweets were included in the analysis. Survivors shared a variety of concerns, with sleep disturbance and weight gain being the most common physical concerns. Popular psychological concerns were depression, body-image, and fear of cancer recurrence, while financial burden of the cancer treatment was a very common social concern among survivors. There was also great variation in the types of support offered by survivors on Twitter. While informational support was the most frequent form of support, instrumental support was the least frequent. Further analysis of informational support revealed that survivors also shared several coping strategies with each other. Engaging in physical activity and spirituality were predominant among these.

Conclusion

Beyond the informational support provided on Twitter, peer survivors also offer positive reinforcement to fellow survivors that they are not alone. Peer support is known to be highly beneficial among cancer survivors. Online platforms such as Twitter can be vital in allowing one to network with fellow survivors, receive needed support, and share coping strategies, especially for those living in remote and/or rural locations.

References

Introduction
The general public’s perception of common health issues such as diet remains limited to survey-based studies. Publicly available Twitter posts have facilitated data collection and leveraged the research at the intersection of public health and data science that cannot otherwise be collected through traditional means of research (e.g., surveys, interviews, focus groups). Furthermore, analyzing Twitter data can help health organizations such as state health departments and large healthcare systems to provide effective health advice and track health opinions of their populations when needed. While previous studies have provided valuable insights, the geopolitical factor of diet has not been investigated on social media. To address this gap in the literature, this study utilizes a computational approach to characterize health-related opinions regarding diet based on the political orientation of the US states. These common perceptions can be used by providers and public health agencies to better understand the common opinions of their population denominators in regard to diet, and reflect upon those opinions accordingly.

Method and Result

Data Collection. We collected 2,348,229 million tweets using Twitter’s Application Programming Interface (API) from June 2016 to May 2017 using diet and #diet. Data Pre-Processing. This step removed duplicate tweets and retweets, stopword, and categorized the tweets based on different states. This process resulted in 857,277 tweets. Topic Analysis. Then we detected the hidden semantic structure using topic modeling in the collected tweets based on the political orientation of the US states.

Decoding Patterns. This phase utilized a qualitative approach to detect, decode, and interpret meaningful patterns with the help of public health experts. This phase removed nine unrelated and meaningless topics and provided 32 topics. Statistical Analysis. Finally, we compared the weight of topics based the political orientation of the US locations including Republican (Rep), Democratic (Dem), and Swing (Swi) states.

Result. We estimated and extracted 41 topics about different issues such as different types of diet, chronic condition, lifestyle, and health behavior. The following table shows that the weight of more than 60% of topics is statistically significant ($P - Value < .05$). Each value in this table shows the percentage of the topics whose weight was statistically significant between each of the categories in the columns. For example, the comparison between the diet topics posted on Republican and Democratic states shows that the average weight of 95.12% of the 41 topics was statistically significant.

Discussion
We found that there was a significant difference between Democratic & Republican, Democratic & Swing, and Republican & Swing states regarding diet-related topics. The key findings as follows. First, the difference of diet-related topics discussed on Twitter between Republican & Swing states is less than Democratic & Swing states. Second, the difference between Democratic & Swing is more than the one between Republican & Swing states and Democratic & Republican. The results are interesting because most Swing states voted for the Republican candidate in the 2016 US presidential election. So, there was a close relationship between Republican and Swing states based on opinions and beliefs.
The Effective, The Inefficient and the Ugly: Interface Elements in Nursing Order Management

Alexandra N. Kassis, Benjamin J. Duncan, MS, Lu Zheng, RN, MHI, MS, David R. Kaufman, PhD, Stephanie K. Furniss, PhD, Adela Grando, PhD, Karl A. Poterack, MD, Richard A. Helmers, MD, Bradley N. Doebbeling, MD, MSc

1 Biomedical Informatics, College of Health Solutions, Arizona State University, AZ, US; 2 Kern Center Informatics and Knowledge Management, Mayo Clinic, Rochester, MN, US; 3 Science of Healthcare Delivery, College of Health Solutions, Arizona State University, AZ, US; 4 Department of Anesthesiology, Mayo Clinic, AZ, US

Description of Problem
Electronic Health Record (EHR) designs that contain unused features put a burden on the user’s cognitive load by cluttering the screen and making navigation complicated. Published reports suggest current EHR design guidelines are not as effective as we need them to be. Investigating the current EHR-mediated workflow in pre-operative nursing order management (OM) at Mayo Clinic allowed us to identify usability problems and frequency of use in selectable features. Identifying the most frequently used functions will aid in the improvement of future OM design while eliminating unused functions will result in workflow optimization, efficiency, better usability and more effective patient care.

Methods
Pre-operative care observations were completed in one of Mayo Clinic’s tertiary hospitals. Usability issues were identified and categorized leveraging a set of 14 usability principles for evaluating health information technology (IT) from Zhang et. al. Problems were identified and categorized by two EHR usability expert informaticists using six of the 14 usability principles for evaluating electronic systems. Selectable features were identified in a system walkthrough and function frequency was calculated using eight recorded and analyzed OM cases. These issues were observed through screen capture using Morae™, a usability video capture software, to record users interacting with the systems in situ. The results presented here were based off observations of eight nurses using SurgiNet in PreOp OM instances at the Mayo Clinic Arizona campus.

Results
Table 1 shows the name of the system in which the usability problem occurs, the specific interface aspect where it occurred, a description of the problem being observed, and which usability principles are involved. There were six observed issues in the Cerner SurgiNet system. The problems with Cerner centered around the display issues including visibility, memory, consistency, and minimalism.

<table>
<thead>
<tr>
<th>System</th>
<th>Interface</th>
<th>Description of Problem</th>
<th>Principle Violated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerner SurgiNet</td>
<td>Order activation/deactivation page</td>
<td>Complex list of all patient orders</td>
<td>Visibility, Memory</td>
</tr>
<tr>
<td></td>
<td>Categorized Order View</td>
<td>Inconsistent Order View vs. Display</td>
<td>Consistency, Visibility</td>
</tr>
<tr>
<td></td>
<td>Order list</td>
<td>Order list with all perioperative orders combined</td>
<td>Minimalism, Visibility</td>
</tr>
<tr>
<td></td>
<td>Order list</td>
<td>Active and inactive orders are combined</td>
<td>Memory</td>
</tr>
<tr>
<td></td>
<td>Orders – main page</td>
<td>Multiple pathways to activate orders</td>
<td>Consistency</td>
</tr>
<tr>
<td></td>
<td>Order category deactivation pop up</td>
<td>Complicated steps to deactivate whole order categories</td>
<td>Minimalism, Memory</td>
</tr>
</tbody>
</table>

Conclusion
The heuristic categorization analysis performed identifies specific usability problems and a description of usability principles violated. Through categorization, usability issues can be identified for further investigation on design improvement. Future work will involve further application of the TURF framework and function frequency on Epic to compare the two-systems interface design and usability. This comparison will allow usability barriers to be more easily identified and suggest future interface aspect redesigns.

References
The Promise and Perils of Pseudonymization

Mehmet Kayaalp, MD, PhD
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

Pseudonymization is the practice of replacing actual personal identifiers with realistic-looking faux identifiers. Its significance has grown immensely as it was adopted by the EU’s General Data Protection Regulation. This study provides a cost-benefit analysis of pseudonymization and concludes that the perils of pseudonymization when used as the default de-identification mode may outweigh the benefits.

Clinical text de-identification systems replace personal identifiers (e.g., James) with appropriate labels such as [Individual-1]. Pseudonymization goes one-step further and maps genuine identifiers to faux identifiers (e.g., James becomes John) within the text and sometimes across the entire collection of data. The resulting de-identified text would read more naturally. Furthermore, if the de-identification system fails to recognize an identifier, leaving it unmodified, the genuine identifier blended among others may remain unnoticeable.1 Due to these benefits, General Data Protection Regulation adopted pseudonymization as the default mode of de-identification.2

Genuine identifiers missed by the de-identification system would be glaringly obvious in the conventionally de-identified output. When one user observes the failure, the policy under which the institution operates should dictate that the incidence be reported immediately, so the underlying problem could be corrected for the entire dataset, preventing further exposure of the confidential information to others. In the same scenario, pseudonymization would only provide an illusion of privacy protection while confidential data could obliviously be distributed to a wider audience. Furthermore, the promise of added privacy protection via pseudonymization has a marginal value since the chance of patient re-identification from the output of a modern de-identification system is usually much less than 1%,3 and when it occurs, the chance that the research scientist knows that patient would be slim to none.

Modern de-identification systems rely on the feedback of their users since the characteristics of input change over time. Pseudonymization, however, may hinder this process in at least two ways: (1) the user may genuinely be unaware of the problem, or (2) without any serious risk of repercussion, the user may pretend to be unaware of the problem to avoid the overhead of reporting and the risk of an extended interruption of their study.

Some pseudonymization algorithms may use one-to-one mapping between actual and faux identifiers across the database for linking patient data longitudinally. If the algorithm or the actual map itself is hacked or revealed, all hidden identifiers could be exposed at once; therefore, such seemingly innocuous methods may pose high risk to privacy and their applications would require great care.

Pseudonymization may also inadvertently alter the actual scientific data by misjudging it as a personal identifier such as a date, age, telephone number, address location, or other numeric identifier. Since pseudonymization creates no sign of alteration, scientists could unsuspectingly use false data, producing erroneous results whose cost would be hard to predict.

In the final analysis, pseudonymization can be appropriate in select circumstances (e.g., sharing well-verified de-identified data to a large group of users); however, using it as the default mode of de-identification may cause more harm than good.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

Representing the Intensity of a Physical Activity

1Yujing Ke, 1Arthur Zhao, 1William Gu, 2Hyeoneui Kim
1Department of Computer Science, 2School of Nursing; Duke University, Durham, NC

Introduction
Assessment and recommendation of a physical activity for a patient needs to specify frequency, duration, types and intensity of the activity. Consistent representations of these physical activity attributes should also be available to draw reliable insights into various health outcomes. Unlike the easily quantifiable attributes like frequency and duration, objectively describing activity intensity is challenging because it is a subjective experience. There exists a few metrics that offer a means to quantifying the intensity level of an activity: Metabolic Equivalent (MET) values can be adjusted to represent the intensity of an activity that a person experiences based on his/her resting metabolic rate1. Borg scale2 also provides a relatively simple but reliable way of assessing the intensity of an activity based on one’s perceived exertion. However, these methods are rarely used in everyday communications of assessment and recommendation of physical activities. This exploratory study investigated a few methods that are widely used to describe and communicate activity intensity information by analyzing standardized physical activity questionnaires.

Methods
We reviewed 1132 unique questions collected from 30 standardized questionnaires developed to assess physical activity habits of adult populations. We manually extracted various expressions describing activity intensity from the collected questions. For example, from a question “Does your work involve vigorous-intensity activity that causes large increases in breathing or heart rate like carrying or lifting heavy loads, digging or construction work for at least 10 minutes continuously?” we extracted three types of intensity descriptors: a general severity modifier “vigorous”, physical exertion “causes large increases in breathing or heart rate”, and activity examples “lifting heavy loads, digging or construction work.” We then analyzed the extracted intensity expressions to discover common patterns and inconsistencies in associating these expressions to different intensity levels of a physical activity.

Results and Discussion
Eighty seven of the 1132 questions contained activity intensity expressions. General severity modifiers were the most widely used method of describing activity intensity and found in 83 questions. We identified 9 severity modifiers such as light, mild, moderate, vigorous, strenuous, heavy, hard, light to moderate, and moderate to vigorous, which were deemed to need further standardization. Activity examples were the next most frequently used method for describing intensity, and were found in 68 questions. Simplified physical exertion levels were used in 36 questions and expressed with changes in heart rate, breathing, and sweating; each was attached with various degree descriptors (e.g., slight increase, faster than normal, soaked, etc.). These three types of descriptions were often used in combination. All three types of descriptions were found in 23 questions like the one provided as an example in the Methods section. A severity modifier was used in combination with activity examples in 41 questions, and with physical exertion in 8 questions. However, more than half of the questions analyzed in this study depended solely on the general severity modifiers to describe activity intensity. The associations made between physical exertion and severity modifiers in these questions were consistent. However, we noted inconsistencies in associating activity examples and severity modifiers to describe activity intensity. For example, bicycling was used as an example of a moderate activity in one question, but presented as a vigorous activity in another. Similarly, swimming and brisk walking were used as examples with both light and moderate intensity.

This study was done in a small scale and not powered to discover various patterns and inconsistencies in intensity descriptions thoroughly. Nonetheless this study suggests that the current methods of describing activity intensity information found in physical activity questionnaires can potentially lead to miscommunication. Describing activity intensity with general severity modifiers, simplified degree of physical exertion, and activity examples might not be a reliable way of conveying intensity information as they all involve in subjectivities in interpretation. This explorative study highlights the need to investigate how people perceive and interpret these commonly used methods of describing activity intensity, and consider adopting a quantifiable method of describing activity intensity.

References
Semantic Representations of Multi-Modal Data, NeuroInformatic Processing Pipelines, and Derived Neuroimaging Results in the Arkansas Image Enterprise System (ARIES)

Aaron S. Kemp, MBA 1,2, Jonathan Bona, PhD 2, Tracy S. Nolan 2, Linda Larson-Prior, PhD 1,2,3, Tuhin Virmani, MD, PhD 3, Lakshmi Pillai, MS 3, Mathias Brochhausen, PhD 2, Lawrence Tarbox, PhD 2, and Fred Prior, PhD 2
1 Psychiatry, University of Arkansas for Medical Sciences (UAMS), Little Rock, AR; 2 Biomedical Informatics, UAMS, Little Rock, AR; 3 Neurology, UAMS, Little Rock, AR

Introduction
Neuroimaging is among the most active research domains for the creation and management of open-access data repositories. Heavy emphasis has been placed on functional magnetic resonance imaging (fMRI) data for both disease specific collections and healthy brain function. The Cancer Imaging Archive (TCIA) has been the National Cancer Institute’s principal imaging resource and has encouraged and supported open-science research by acquiring, curating, hosting and managing collections of multi-modal information. The TCIA technology stack is currently being refactored into a more streamlined, easily maintained, containerized package which we have labeled PRISM: Platform for Imaging in Precision Medicine. One of the first applications of PRISM is the establishment of a neuroimaging research data management system at the University of Arkansas for Medical Sciences, which is known as the Arkansas Image Enterprise System (ARIES). As the first instantiation of the PRISM infrastructure, the ARIES project aims to explore the practical utility and usability of the full set of capabilities that this new platform provides. In particular, the integration of semantic representations of multi-modal data elements from a variety of disparate sources (e.g., imaging, behavioral, or cognitive assessments), across image processing stages (e.g., preprocessing steps, neuroanatomical segmentation schemes, analytic pipelines), as well as descriptions of the derived results would ensure greater reproducibility and comparability of scientific findings across large-scale neuroimaging research projects.

User Groups and Pilot Data
Pilot testing of the ARIES instantiation of PRISM is being conducted with three collaborating investigative teams who are using ARIES in a project designed to identify common pathways of neurodegeneration. The dataset for the pilot test includes neuroimaging measures (structural and functional MRI and EEG) as well as endophenotypic data obtained from a variety of assessments designed to measure neuro-motor integration (wearable body sensors, gait-assessment floor mat, digitized gloves, and handwriting/drawing assessments on a digitizing tablet) and neurocognitive functions (performance scores on standardized neuropsychological tests and cognitive activation tasks from functional imaging) in three unique study cohorts diagnosed with Parkinson’s disease (PD), Mild Cognitive Impairment (MCI), or Cancer-Related Cognitive Impairment (CRCI).

Semantic Integration Approach
To integrate and manage these data we are building semantic representations using axiomatically-rich ontologies. These will be used to instantiate a knowledge graph that combines the data from these unique study cohorts into a shared semantic representation that explicitly accounts for relations among these data. This amounts to providing queryable relationships across the source data sets. This knowledge graph is stored in a triple store database that supports reasoning over and querying these integrated data. We believe that these unique capabilities will facilitate the discovery of important new linkages among endophenotypic expressions of disturbed neural functions and discrete neuroanatomical markers of neurodegeneration obtained from the derived neuroimaging results.

Conclusion
Semantic integration of neuroinformatic processes in the ARIES pilot project demonstrates the capabilities of the PRISM infrastructure to effectively represent detailed neuroanatomical segmentation schemata and processing pipelines for image analyses, integrate a diverse set of multi-modal data elements, and provide detailed descriptions of the results obtained across the analytic processing stages and in relation to the supporting endophenotypic data. Such capabilities are essential to ensure greater reproducibility in large-scale neuroimaging research projects.

References
Sickle Cell Disease Phenotype Algorithm Performance in Adult, Pediatric, and Transitional Care

Mirza S. Khan, MD\textsuperscript{1,2}, Mark Rodeghier, PhD\textsuperscript{3}, Robert M. Cronin, MD, MS\textsuperscript{2}
\textsuperscript{1}U.S. Department of Veterans Affairs, Nashville, TN; \textsuperscript{2}Dept of Biomedical Informatics, Vanderbilt University, Nashville, TN; \textsuperscript{3}Rodeghier Consultants, Chicago, IL

Introduction
Sickle cell disease (SCD) is an inherited condition that is associated with increased healthcare utilization and mortality. This risk is greatest in adolescent patients due to the poor transition of care from pediatric to adult-based practice\textsuperscript{1}. Adherence to evidence-based SCD management is sub-optimal. To prevent complications, efficient methods are needed to monitor care practices in this population. Computable phenotype algorithms are a promising method to query electronic health record (EHR) data to discover patients with rare conditions, such as SCD.

Methods
We conducted a retrospective observational study using EHR data from 2012 to 2016 at Vanderbilt University Medical Center and Monroe Carell Jr. Children’s Hospital at Vanderbilt. This study was approved by the Vanderbilt University IRB. Similar to prior work, algorithm 1 queries pertinent ICD-9 and ICD-10 codes\textsuperscript{2}. Algorithm 2 uses relevant SNOMED terms and word matching. Algorithm 3 identifies cases satisfying both algorithms 1 and 2. We used clinical encounter metadata to categorize patients by clinical setting as pediatric, adult, and transition. The latter accounts for patients transitioning from pediatric to adult practice. SCD status was confirmed using a manually curated SCD patient registry. We calculated positive predictive value (PPV) and sensitivity with and without stratification by age group. We used the \(\chi^2\)-test to study SCD status between age categories and the Fisher’s exact test to compare PPV and sensitivity by age group for algorithms 1 and 2. Statistical analysis was performed using SPSS version 25.

Results
The study included 728 total patients; 87.2\% were black and 50.4\% were female. PPV ranged from 88.4-94.0\% and sensitivity from 89.2-100\% (Table 1). True SCD status between age groups is significantly different (\(p = 0.002\)). Sensitivity between age groups is different for Algorithm 1 (\(p = 0.009\)) and for Algorithm 3 (\(p = 0.034\)). There is no difference in PPV (\(p = 0.454\)) and sensitivity (\(p = 0.136\)) between Algorithms 1 and 2.

Table 1: Summary statistics, PPV and sensitivity calculated by age group for Algorithms 1, 2 and 3.

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample size</th>
<th>No. true SCD</th>
<th>Mean age (yrs)</th>
<th>Algorithm 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Algorithm 2</th>
<th></th>
<th></th>
<th></th>
<th>Algorithm 3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pediatric</td>
<td>344</td>
<td>321 (93.3%)</td>
<td>6.4</td>
<td>PPV (%)</td>
<td>Sensitivity (%)</td>
<td>PPV (%)</td>
<td>Sensitivity (%)</td>
<td>PPV (%)</td>
<td>Sensitivity (%)</td>
<td>PPV (%)</td>
<td>Sensitivity (%)</td>
<td>PPV (%)</td>
<td>Sensitivity (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transition</td>
<td>37</td>
<td>34 (91.9%)</td>
<td>17.6</td>
<td>93.4</td>
<td>96.9</td>
<td>93.8</td>
<td>94.1</td>
<td>94.0</td>
<td>93.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>347</td>
<td>295 (85.0%)</td>
<td>34.0</td>
<td>91.2</td>
<td>91.9</td>
<td>88.4</td>
<td>90.5</td>
<td>92.6</td>
<td>89.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>728</td>
<td>650 (89.3%)</td>
<td>20.1</td>
<td>92.4</td>
<td>94.8</td>
<td>91.2</td>
<td>92.8</td>
<td>93.3</td>
<td>91.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion
Computable phenotype algorithms are efficient, scalable and cost-effective alternatives to manual patient registries for identifying patients with rare conditions. Using SNOMED terms and word matching resulted in comparable performance to our algorithm that queried ICD codes. Furthermore, the PPV for each algorithm is not different across age groups, even among those undergoing care transition. Use of such algorithms to identify patients with rare diseases may aid population health monitoring and provides efficient cohort development for research studies.

References
A Review of Palliative Care Mobile Applications for Patients and Caregivers

Eugenia Kim, MS1,2, Fariha Ahsan, MPA1,3, Corinne Lamour-Romero, RN1,2, Chelsea Michael1,2, Ruth Masterson Creber, PhD, MSc, RN1
1Weill Cornell Graduate School of Medical Sciences, New York, NY; 2Memorial Sloan Kettering Cancer Center, New York, NY; 3NewYork-Presbyterian Hospital, New York, NY

Introduction

As the elderly population continues to grow, understanding how mobile technology can be used to improve end-of-life and palliative care for aging patients and to support their caregivers is becoming increasingly important. Within the United States, older people (ages 60 and over) are expected to outnumber children and young adults by 2030.1 While research shows that smartphone use and ownership varies by age, utilization of smartphones has grown most dramatically in the 60 and older age groups.2 However, to our knowledge, no studies have systematically evaluated consumer apps for palliative care. We reviewed mobile apps designed to improve palliative care of patients and support their caregivers.

Methods

In February 2019, a comprehensive review of mobile apps was conducted across three mobile app stores: Apple iTunes store using fnd.io, Android Google Play store, and Amazon Appstore. The following search terms were included: “palliative care,” “end-of-life care,” “hospice,” “hospice care,” “hospice caregiver,” “hospice support,” and “terminal illness.” Apps were evaluated using the Mobile Application Rating Scale (MARS), IMS Institute for Healthcare Informatics functionality assessment, and evidence-based guidelines for self-management based on the Supportive and Palliative Care Indicators Tool (SPICT) and Palliative Care Network of Wisconsin (PCNOW) guidelines, which included criteria such as symptom management and enhanced communication.

Results

We identified 1856 apps, of which 19 met the final inclusion criteria. Most applications were excluded because they were non-English, games/books, unrelated, for clinicians only, or medical information and education. The most common criteria that was addressed from the SPICT/PCNOW guidelines were reviewing plan of care and easing mental health suffering. Interrater reliability among four reviewers was substantial (0.58-0.79). The majority of apps did not meet the criteria recommended by the SPICT/PCNOW guidelines. The median MARS score was 2.6 out of 5, and eight of the 19 apps met the minimum acceptability score of 3.0. Across all our reviewed apps, the average IMS functionality score was 3.3 out of 11 with the most common features being inform (20%) and communicate (20%) (Figure 1). Overall, the Mind Your Loved Ones app scored highest on the MARS, IMS functionality, and SPICT/PCNOW guidelines. The most common behavioral change approach was to “Inform/Educate” (11 apps). Ten apps required web access to function and seven allowed sharing information from the app to loved ones or to popular social media platforms. Six required logins, but none offered data encryption.

Discussion & Conclusion

This review indicates that of the currently available apps, few meet the identified criteria for quality or functionality. Many of the palliative care apps were not grounded in theory or evidence-based guidelines. In particular, despite an emphasis on recommendations for symptom management in palliative care, there was a noted absence in the apps providing any such features to support patients and caregivers in doing so. These findings suggest that apps need further targeted development, and policy makers, app developers, and designers should work collectively to create both policies and apps that are founded on evidence-based clinical guidelines.

References

Developing computational phenotyping algorithm to classify lung cancer stages using common data elements

Min-Hyung Kim, MD, MS, Sajung Park, MD, Yu Rang Park, PhD, Wonjun Ji, MD, Seul-gi Kim, MA, Minji Choo, RN, Seung-sik Hwang, MD, PhD, Jae Cheol Lee, MD, PhD, Hyeong Ryul Kim, MD, PhD, Chang-Min Choi, MD, PhD

1 School of Public Health, Harvard University, Boston, MA, USA
2 College of Medicine, University of Ulsan, Asan Medical Center, Seoul, Republic of Korea
3 College of Medicine, Yonsei University, Seoul, Republic of Korea
4 Graduate School of Public Health, Seoul National University, Seoul, Republic of Korea

Introduction

TNM stage is an important classification system for predicting prognosis and planning treatments in lung cancer. Unfortunately, TNM information is not available in ICD-9/10 systems, and this limits analyses of large-scale health databases. We aimed to develop computable phenotype algorithms to classify the TNM cancer stages of lung cancer patients using common data elements, including diagnosis, medication and procedure codes.

Methods

We first developed a rule-based computational phenotyping algorithm based on medication and procedure codes, according to the recommended treatment regimens in the U.S. National Comprehensive Cancer Network (NCCN) guidelines. Then we validated the base classifier in electronic health records of a tertiary referral hospital. The gold-standard staging information was attained by manual chart review for 2,907 lung cancer patients. We divided the patients into three random subsets and used the first subset for the initial validation. Then, we developed three additional rules: the second rule added limits based on prior non-pulmonary malignancy, the third rule added limits based on pulmonary comorbid conditions (ICD-10 A15-16, J10-18, J41-47, J84-85), and the fourth rule added limits based on both prior non-pulmonary malignancy and pulmonary comorbid conditions. The performances of all algorithms were evaluated in the second subset, and the third subset remains sequestered for future research.

Results

After reviewing the initial validation results (overall accuracy 0.75, pooled sensitivity 0.63, specificity 0.91, c-statistics 0.77), we developed the additional rules and validated the classification performance of each rule in addition to the base classifier in the second subset. In the secondary validation, the base classifier showed overall accuracy, pooled sensitivity, specificity and c-statistics of 0.77, 0.62, 0.92, and 0.77, respectively. When the second rule was added to the base classifier, the overall accuracy, pooled sensitivity, specificity and c-statistics were 0.79, 0.62, 0.91, and 0.77, respectively. When the third rule was added to the base classifier, the overall accuracy, pooled sensitivity, specificity and c-statistics were 0.78, 0.63, 0.92, and 0.78, respectively. When the fourth limit was added to the base classifier, the overall accuracy, pooled sensitivity, specificity and c-statistics of 0.79, 0.62, 0.91, and 0.77, respectively.

Discussion

In this study, we show validation results of rule-based computational phenotyping algorithm based on common data elements, including medication and procedure codes, which can be readily available in large-scale health databases. The phenotyping algorithm based on treatment patterns may have high sensitivities and specificities when there is an authentic clinical guideline that clinicians strictly follow. On the other hand, there are certain scenarios when clinicians make exceptional decisions in practice. This study highlights the opportunities of challenges in developing computational phenotyping algorithm to classify lung cancer stages using common data elements.

References

2. World Health Organization. ICD-10 [Internet]. Available from: http://apps.who.int/classifications/icd10
The Effectiveness of using mobile EMR with consultation process in EMR

Su Jin Kim¹, BA, Kwang Yul Jung, MD², Won Chul Cha, MD³

¹ Department of Digital Health, SAIHST, Sungkyunkwan University, Seoul, Korea
²Department of Emergency Medicine, Inha University School of Medicine, Incheon, Korea
³Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

Introduction

Consultation is critical part of Emergency Department (ED) function.[1] Delayed response from specialty physician results in disposition delay and ultimately attributed to ED crowding which has been critical concern to public health at the aspect of throughput process of ED crowding stages. Mobile electronic medical records (mEMR) is increasingly utilized by health care providers.[2] Several studies has proven that mEMR makes physician work efficiently. However, none of previous studies have evaluated the time efficiency in real clinical circumstances depending on the use of mEMR. In this study, we identified the relation between consultation response and frequency of mEMR.

Method

This retrospective study took place in an academic ED in Seoul, South Korea. We analyzed PC EMR and mEMR data from May 2018 to December 2018. From PC EMR, Request time for consultation from ED physician and record written time by other specialty physician were extracted. Frequency are extracted from log data of mEMR. We merged two data by physician ID and performed Pearson’s product-moment correlation to identify the association.

Result

During the study period, the total of 40,991 cases of consultant from ED physician occurred. 277 physicians received request of consultation. From the 277 physicians, 216 were mEMR users. We found that the frequent mobile EMR users respond more quickly to consultation from EMR. (coefficient: -0.25(-0.37 to -0.17), P-value< 0.001)

Conclusion

This study support that the use of mEMR can enhance the time efficiency in real clinical circumstances.

References

Cancer Type Classification by Jointly Using Words and Concepts from Electronic Health Record Text Notes

Youngjun Kim, PhD1, Stephane M. Meystre, MD, PhD1

1 Medical University of South Carolina, Charleston, South Carolina, USA

Introduction: Electronic health records (EHRs) contain a wide variety of data important to patient care management. The vector representation of this medical information has been increasingly applied to clinical text classification problems. This study investigates what information contained in the EHR contributes to classifying patient cancer types. Our task involves recognizing three types of cancer: breast cancer, gastrointestinal cancer, and lung cancer. The classifier assigns one of the three cancer categories to an EHR using its textual content. We show that document classification can be improved when the words and medical concepts contained in EHRs are jointly used. We also propose a method to extract synonyms from the medical knowledge-base by comparing two concepts that share the same meaning. The synonym list is used to refine the word vector embeddings (WV) for more accurate classification.

Methods: We used EHR text notes from 154 randomly selected cancer patients treated at the Huntsman Cancer Institute (Salt Lake City, Utah). We selected the 168 most frequently diagnosed medical problem at the University of Utah Healthcare in 2013. For each note, all mentions of these problems were annotated by medical experts with about 70% inter-annotator agreement. The corpus consisted of 495 notes used for training with 9,337 problems and 274 notes used for testing with 5,360 problems. To automatically identify notes from patients with breast, gastrointestinal or lung cancer, we created four different multi-class classifiers for document classification. We used the Vowpal Wabbit online learning library with 3-skip-trigrams that allow words to be skipped for each word in the document. One-vs-all classification strategy was applied to perform multi-class classification. We used fastText for the other three classifiers: one without pre-trained WV, one with pre-trained WV, and one with refined WV using concept synonyms from the UMLS Metathesaurus. We used all terms associated with UMLS concepts and created pairs of concepts. For each concept term in the pair, we removed common words that occurred in both terms. When only one word remained in each of the two terms, we considered them as synonyms. We rebuilt the WV representations by applying the refinement method proposed by Faruqui et al. with this list of synonyms. For all four classification methods, we implemented three different variations: using only words, only concepts, and both words and concepts.

Results: We report the accuracy of each model correctly identifying the type of cancer in each patient note. The fastText models showed higher accuracy than the corresponding Vowpal Wabbit models. Pre-trained word embeddings contributed to the accuracy when words were used for training. The fastText model, with refined word embeddings using synonyms produced the best accuracy at 93.5% when both words and concepts were used as features. The paired t-tests showed that the accuracy of this model was significantly better than models preceded by an asterisk (*) in Table at the p < 0.05 significance level.

Conclusion: Instead of using only diagnostic codes or only words found in EHR text notes, we demonstrated that jointly using words and standard terminology concepts can further improve document classification for identification of patients with specific cancer types. The automated extraction of synonyms from the standard terminology yielded promising results. Yet our study only covered one-word synonym terms. One direction for future research involves comparing pairs of synonymous terms with multiple words.

Acknowledgments: This research was supported in part by the National Cancer Institute (R41CA180190).

References

1632
Insights from a dissertation on the development of a Learning Electronic Medical Record System: data-driven, context-aware learning

Andrew J. King, PhD\textsuperscript{1,2}, Shyam Visweswaran, MD PhD\textsuperscript{1,3}, Harry Hochheiser, PhD\textsuperscript{1,3}, Gilles Clermont, MD MSc,\textsuperscript{2} Gregory F. Cooper, MD PhD\textsuperscript{1,3}

\textsuperscript{1}Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
\textsuperscript{2}Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, USA
\textsuperscript{3}Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

The vast amount of data stored in a patient’s Electronic Medical Record (EMR) makes it difficult for clinicians to get the information they need. We use machine learning to predict which patient data a clinician is likely to seek, and we highlight those data in the EMR. In this poster, we report insights gained while developing and evaluating a Learning EMR system.

Methods

A Learning EMR observes clinician information-seeking behavior and applies it using machine learning to direct the display of data in future patient cases. Over the last five years, we have developed, trained, and evaluated a Learning EMR system. In the most recent work, we evaluated a Learning EMR system that was trained on the information-seeking behavior of 11 intensive care unit (ICU) clinicians (intensivists) when, in a laboratory environment using de-identified EMR data, they prepared to present about 200 patient cases at morning rounds. The evaluation consisted of 12 intensivists each performing the same clinical tasks for 18 patient cases, where the cases were presented using either the control “vanilla” EMR user interface or using an intervention “highlighted” EMR user interface. The experience of conducting this work culminates into insights into context-aware systems such as the Learning EMR.

Results

This poster presents insights into \textit{what}, \textit{where}, and \textit{when} data should be highlighted in a Learning EMR, and \textit{how} training data should be collected for a Learning EMR.

Conclusion

Data-driven approaches for adaptive EMR systems show promise for supporting clinical decision making and enhancing user experience with EMRs. These insights will help others pursuing similar problems involving context-aware learning.

References and Acknowledgements


- NLM T15 LM007059; R01 LM012095; R01 GM088224
Continuous Predictive Modelling for Sepsis, Organ Failure, and In-hospital Mortality in the Intensive Care Unit

Amber Kiser, BS¹, Devin Horton, MD², Karen Eilbeck, MSc, PhD¹, Samir Abdelrahman, MSc, PhD¹
¹Department of Biomedical Informatics, University of Utah, School of Medicine, Salt Lake City, Utah
²Department of Internal Medicine, Division of General Internal Medicine, University of Utah Medical Center, Salt Lake City, Utah

Introduction: Patients in the Intensive Care Unit (ICU) can deteriorate very quickly, leading to unexpected outcomes, including sepsis, organ failure, and death. Sepsis is a dysregulated host response to infection, which for this project is identified clinically as systemic inflammatory response syndrome (SIRS) plus infection. Most current clinical risk scores such as SOFA are calculated once a day. Because of this, there is little ability to know how things are changing in real time. Additionally, previous studies have demonstrated that clinical risk scores such as SIRS and qSOFA have lower positive predictive value (PPV), despite having acceptable area under the receiver operating characteristics (AUROC). Continuous predictive modelling, however, can identify negative trends early and may improve patient management and outcomes. The goal of this project is to develop continuous predictive modelling for each outcome, i.e., sepsis, organ failure, and in-hospital mortality, in order to monitor the patient’s status hourly. Several recent models have been developed to predict sepsis with at least one lacking precision when compared to the qSOFA score until 6 hours of data were incorporated into the model, demonstrating a need for continuous modelling. Our continuous modelling predicts sepsis along with sepsis-related outcomes, organ failure and mortality, using only 1 hour of data with greater precision when compared to the risk scores most appropriately matched for each outcome.

Methods: The data was retrieved from the MIMIC III database. Patients under the age of 18 were excluded. Only patients who were transferred to the ICU at some point during their hospital stay were included. In total, 38,439 unique patients were included, comprising 8,301,364 rows of hourly data. Vital signs and laboratory results were binned hourly for each patient. For each hour, the outcomes were determined from the vital signs and laboratory values, including whether the patient had sepsis, was in organ failure, or if the patient died. The final cohorts were prepared individually for each outcome. Vital signs, laboratory values, patient age, and some medication information were used as features in the model. In order to avoid biasing the models, some features were excluded from certain models because these measurements were used in defining the outcome. Observations from the previous hour were used to predict outcomes for the current hour. Several clinical risk scoring criteria were used to compare the PPV and NPV of the continuous prediction models. These included SIRS, used to compare with sepsis prediction, qSOFA, used to compare with organ failure prediction, and mEWS and NEWS2, used to compare with mortality prediction. Each hour, these clinical risk scores were calculated for each patient based on the published criteria and standard thresholds. The datasets were split into training and testing sets in an 80/20 split by hourly records. The training data was used to perform minimal hyperparameter tuning for 5 different algorithms, including Gradient Boosting, SVM, Logistic Regression, Naïve Bayes, and Random Forest. After training, the models were evaluated on the unseen test data and compared to the risk scores. Our hypothesis was that the continuous predictive models would have greater precision than their matched risk scores.

Results: For all outcomes, logistic regression underperformed against the clinical scores. However, Naïve Bayes did perform better than the scores in sepsis prediction and SVM performed better than the scores for organ failure prediction. For sepsis prediction, the gradient boosting (PPV = 0.348, NPV = 0.989, AUC = 0.877) and random forest (PPV = 0.358, NPV = 0.990, AUC = 0.961) models produced comparable results. Both surpassing the SIRS risk score (PPV = 0.001, NPV = 0.985). For organ failure prediction, the gradient boosting model (PPV = 0.266, NPV = 0.980, AUC = 0.754) surpassed the qSOFA risk score (PPV = 0.054, NPV = 0.980) in PPV and produced a comparable NPV. Finally, for mortality prediction, once again the gradient boosting model (PPV = 0.196, NPV = 0.999, AUC = 0.836) surpassed the mEWS (PPV = 0.006, NPV = 0.999) and NEWS2 (PPV = 0.002, NPV = 0.999) risk scores in PPV while again producing comparable NPV.

Discussion: This study served as a proof of concept for our hypothesis that continuous modelling would predict sepsis and sepsis-related outcomes better than clinical risk scores. Our hypothesis was demonstrated to be true as the top performing continuous prediction models resulted in higher PPV than the risk scores associated with the outcome. There are some limitations in comparing different study models which predict sepsis. For example, these studies often use different definitions of sepsis and predict different time periods prior to onset. However, given this caveat, our model performed better or comparable to four recent models, granted they are predicting different time periods before onset. Our best performing random forest model produced an AUC = 0.961, compared to Delahanty et al AUC = 0.96, Desautels et al AUC = 0.88, Nemati et al AUC = 0.85, and Taylor et al AUC = 0.860. In order to more accurately and appropriately compare the current study model to other similar models, we plan to incorporate the ability to predict sepsis onset more than one hour in advance, using data from more than just the previous hour. Additional future enhancements to our study include different validation methods, such as cross validation and a validation dataset, different feature selection methods, additional datasets including different patient populations, and more advanced techniques such as deep learning.

References:
Virtual Reality: Exploring Methods to Improve Dietary Choices

William Kistler, MA1,2, Denise M. Goldsmith MPH, MS, RN, FAAN1, Patricia Flatley Brennan PhD, RN1,3
1National Institute of Nursing Research, Washington, DC; 2University College London, London, UK; 3National Library of Medicine, Washington, DC.

Introduction

Virtual Reality (VR) is showing promise in healthcare1 and may be of value in rehearsing self-management tasks for patients with complex disease such as heart failure (HF). Patients with HF experience difficult life-style alterations, including the need to follow a sodium restricted diet.2 As part of the development of an advanced visualization laboratory to investigate patient self-management, we conducted focus groups with experienced home care nurses to identify opportunities for patient self-care that could be investigated using VR. In this effort, we use the results of the focus group to explore a proof of concept VR environment wherein patients can select food according to their sodium restricted diet and receive feedback on their choices.

Methods

We invited four experienced home care nurses to participate in a series of two, ninety-minute focus groups designed to identify self-management concerns for chronically ill patients. Food choice selection was identified as a common challenge for patients with HF on a restricted dietary sodium regimen. To investigate whether a VR environment could facilitate food selection behaviors for patients on a sodium restricted diet, we drew storyboards to depict individuals confronted with a range of high and low sodium food choices. The use of VR to replicate and facilitate this type of real-world food choice selection process has been validated in earlier research, and we will leverage that work.3 This environment will be built using Blender 3D and Unity3D engines for use with the HTC Vive™.

Discussion

Proof of Concept Design Considerations: we chose to develop a shopping experience in a grocery store environment within VR store ex in VR that will allow the user to select food for purchase, and review choices. Periodically, the user’s selected foods will be calculated for sodium levels and feedback will be provided to the user. Users will be able to exchange foods or continue with their current selections. Our user experience and VR environment will expand on prior work in its use of feedback to user choices and monitoring user responses. Using VR enables experimental and experiential benefits. Of note, all VR participants are exposed to an identical visual and auditory experience. As such, investigators are given complete control of the environment and its experimental variables. VR environments can also provide quantifiable data for behavioral analysis, i.e. every food choice and body movement could be calculated and recorded in a way that permits detailed analysis. There are notable tradeoffs to using a virtual environment. It is relevant to note that there would likely be no sense of smell or touch available in our VR environment. These senses would be valuable for a food selection task; nonetheless, similar research in diet and food choice has shown success without requiring these sensory stimulations.3 Another barrier to our proposed virtual environment arises from the limited resolution of VR displays. Currently, smaller text sizes are difficult to read, even for individuals with normal vision. To effectively navigate this challenge, another form of interface with the nutrition information may be required, which may in turn reduce the veridical experience for the user. Future studies conducted in this environment may support the clinician’s ability to identify whether a patient understands their prescribed sodium restrictive diet and determine which mode of feedback to food choice could best prevent sodium overload or hospital readmission.

References

Detecting and mitigating clinical guideline interactions in multimorbidity patients using computer-interpretable guidelines and AEOLUS

Alexandra Kogan¹, BSc; Mor Peleg¹, PhD; Samson W Tu², MSc

¹Department of Information systems, University of Haifa, Israel
²Stanford Center for Biomedical Research, Stanford University, Stanford, CA, USA

Introduction. Clinical practice guidelines (CPGs) provide evidence-based recommendations for single diseases while addressing prevalent comorbidities and side effects. In previous work, we developed a decision-support system with a Controller component that can detect and mitigate interactions among computer-interpretable guidelines (CIGs) by reasoning on the CIG knowledge using predefined interaction-management design patterns. Our goal-based method utilizes knowledge of drugs’ physiological effects and therapeutic usage to combine knowledge from CIGs. It incrementally detects interactions and plans non-contradicting therapies. The Controller algorithm [1] uses patterns to check consistency and respond to events such as data enquiries, diagnoses, adverse events, recommended medications, tests, and goals. The method utilizes existing standards and CIG tools, including the Fast Healthcare Interoperability Resources (FHIR) patient data model, National Drug File - Reference Terminology (NDF-RT), and the PROforma CIG formalism with its Alium knowledge-engineering environment and PROforma enactment engine. In this work we expand the sources of knowledge used by the Controller to include, in addition to CIGs and NDF-RT, a repository of drugs’ side effects. We demonstrate the use of a design pattern that interacts with the side effect repository using a case study of a diabetes Mellitus (DM) medication causing a recurrent Fungal Infection (rFI).

Methods. In this study, we linked the Controller to the Adverse Event Open Learning through Universal Standardization (AEOLUS) [2]. This resource contains standardized reports and pre-computed statistics derived from the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS). AEOLUS contains information about the Proportional Reporting Ratio (PRR). We defined a drug-disease-interaction pattern that directs the Controller to search for any drug-disease interactions encoded in relevant CIGs, and if no interactions are found, send a query to the AEOLUS database with the RxNorm ID numbers of each of the medications that the patient is currently taking. The query returns possible side effects for each of the medications. After consulting with our Medical expert, we decided to include side effects that have a PRR>2. The Controller compares the side effects returned with the name and ID of the new diagnosis and presents the user with the medications that could be the causes or contributing factors of the new condition, as well as the PRR statistics. In addition, an Adverse Event FHIR object is logged in the system, with the relevant medication and condition.

To demonstrate the Controller’s reasoning with side effects and conclusion of adverse events [3], we represented in Alium the DM and rFI CPGs and analyzed a case of a patient with DM taking Metformin and Gliflozin (SGLT2 Inhibitor), who develops rFI. The CPGs do not reference SGLT2I as a potential contributing factor of rFI.

Results

![Figure 1. The Controller’s dialogue for a patient with DM (Metformin+Gliflozin) who develops rFI showing detection of fungal infection as an adverse event in the AEOLUS knowledge base](image)

Discussion. In this study we demonstrated the behavior of a Controller that can detect and mitigate drug-disease interactions, using a case study involving two morbidities (DM and rFI), where the second morbidity is a potential adverse event of a medication prescribed to treat the first morbidity. The system that was developed based on [1] was evaluated with ten 6th year medical students and was well received. The detection of the potential adverse event supplies the physician with additional information that is not present in the guidelines and might be missed otherwise. This allows for a better treatment for the patient.

Acknowledgement. This project has received funding from Israel Science Foundation grant 906/16.


Chronic Condition Symptom Enrichment from Electronic Health Records

Theresa A. Koleck, PhD, RN1, Suzanne Bakken, PhD, RN, FAAN, FACMI1, and Nicholas P. Tatonetti, PhD1
1Columbia University, New York, NY, USA

Introduction
Symptoms are a significant source of burden for patients living with chronic conditions, but remain understudied and poorly understood. Electronic health records (EHRs) offer an opportunity for longitudinal data-driven characterization of symptoms after chronic condition diagnosis. Our previous efforts to analyze symptoms of chronic conditions using structured EHR data have proved largely uninformative due to the over influence of symptom documentation related to emergency room (ER) visits (e.g., chest pain, dizziness) rather than the condition itself (e.g., fatigue). Therefore, the purpose of this work is to evaluate symptom enrichment in four common chronic conditions – chronic obstructive pulmonary disease (COPD), heart failure (HF), type 2 diabetes mellitus (T2DM), and cancer – via normalization.

Methods
We identified patients with ≥ 2 International Classification of Diseases, 9th and 10th Revisions (ICD-9/10), diagnosis codes on different days for COPD (490-492, 494, 496; J44), HF (428; I50), T2DM (250; E11), or cancer (140-209; C00-97), using structured inpatient and outpatient EHR data at Columbia University Medical Center from 2007 to 2017. For these patients, we extracted ICD-9/10 codes for 80 symptoms (780-799; R00-69) occurring after the initial date of condition diagnosis. We calculated the percentage of distinct patients with ≥ 1 code for each symptom by condition. We then generated 20 datasets of 50,000 randomly selected patients from the entire EHR population with ≥ 2 codes documented between 2007 and 2017. We used the mean and standard deviation of the percentage of distinct patients with ≥ 1 code from the random datasets to generate z-scores for each symptom. We compared top symptoms for the four chronic conditions before and after normalizing by population-level symptom documentation.

Results
The sample sizes and percentage of patients with ≥ 1 symptom documented for COPD, HF, T2DM, and cancer were n=18,219, 75.1%; n=29,569, 74.4%; n=52,837, 76.1%; and n=53,392, 53.0%, respectively. The mean percentage of patients with ≥ 1 symptom documented from the random datasets was 29.3%. The number of symptoms out of 80 that occurred at a significantly greater (Bonferroni corrected p<0.00016) percentage in the condition than in the population was 60, 57, 61, and 56 for COPD, HF, T2DM, and cancer, respectively. Table 1 shows the top 5 symptoms based on percentage before normalization and based on z-score after normalization.

Discussion
We determined enriched symptoms for COPD, HF, T2DM, and cancer after diagnosis from structured EHR codes by normalizing using population-level EHR symptom documentation. The percentages of patients with symptoms documented were higher for each chronic condition both overall and for the majority of individual symptoms compared to the randomly selected datasets. Normalization resulted in emergence of symptoms not traditionally associated with ER visits, including debility and nocturia. Of particular interest is the transition from chest pain to other malaise and fatigue as the top enriched symptom for cancer. While the magnitudes of the z-scores suggest the need for additional refinement of controls, normalization will be an essential step of our future symptom characterization analyses. Funded by K99NR017651 and P30NR01658.
A Review of Older Adults’ Personal Health Information Management Processes, Needs, Challenges, and Tools

Malgorzata Kolotylo-Kulkarni, PhD Candidate¹, Cynthia M. LeRouge, PhD¹, Deborah E. Seale, PhD²

¹Florida International University, Miami, FL; ²Des Moines University, Des Moines, IA

Introduction & Background. Patient access to, and effective management of their personal health information (PHI) are critical in supporting patients in their health management. Due to their growing healthcare needs, effective PHI management (PHIM) may be particularly helpful for older adults. In order to ensure effective PHIM and to provide functional PHIM technology, we need to have a thorough and comprehensive understanding of the idiosyncratic PHIM practices by the respective target groups. It is therefore critical to establish a comprehensive view of their practices and needs to examine whether older adults differ in their PHIM practices, needs and preferences from other age groups.

To tailor PHIM technology to an aged population, it is also important to explore whether the differences in older adults subgroups with regard to their healthcare needs and practices (for instance due to life transitions – retirement and associated with it insurance changes, caregiving – or adequately, requiring a caregiver) materially affect PHIM needs, challenges, and practices, and the implications on technology design. Notable subgroups include 1) consumers aged circa 50 to 69 years old (including baby boomers); and 2) the elderly - consumers of age 69 and above (presently The Silent Generation and The Greatest Generation).

While a few reviews of PHIM literature exist, not many have focused on older adults, and more so, none of them have provided an overall view of the various factors which play a role in PHIM. Literature related to older adults remains fragmented and does not provide a comprehensive understanding of PHIM among this subpopulation and its nuances. This study aims to address the need to better understand PHIM for older adults through a systematic literature review covering current empirical evidence of the practices and tools which older adults use for PHIM, their PHIM needs, and the challenges they face. We will address this purpose by answering the following research questions:

• What idiosyncratic PHIM attributes for an older adult subpopulation is revealed by the literature?
• Do the aspects related to PHIM differ across the two recognized age-related subgroups among older adults?

We adopt the Patient Work Model¹, which references 1) person-related factors; 2) socio-organizational aspects; 3) tasks carried out; 4) physical environment; and 5) tools used and information managed.² as comprehensive lens to organize extant findings and to establish current knowledge on the elements of PHIM by older adults.

Methodology. This study carries out a review of literature from 2008 to 2018, adopting adopt the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) model. The literature search process in our study consists of 2 stages: 1) systematic search using online databases; and 2) backwards and forward citation analysis of the full papers found during the search and subject to eligibility analysis. We have so far completed the first stage of this process, and have run a systematic search in 13 databases, with a keyword search of “health information management” OR “personal health information management”; OR “consumer health information management” in papers’ titles or abstracts. Inclusionary criteria incorporate: empirical papers published between 2008 and 2018, written in English, full-text available, peer reviewed, conference papers and abstracts, address one or both age subgroups, and examine one of the constructs in the Patient Work Model. Analysis of the literature is being carried out using qualitative methods according to the Patient Work Model and using Dedoose©.

Early Research Findings. At present stage of the study our review has identified n=15 papers. The final sample of articles is anticipated to provide a comprehensive view of the empirical evidence on PHIM by older adults and systematize it by the elements of older adults’ patient work system³. Preliminary findings indicate that PHIM by this population is highly complex in terms of the various environmental aspects: 1) older adults face particular challenges (e.g. having experiential health knowledge may impact their health information needs); 2) social and organizational support (such as training/education) are important for older consumers; 3) PHIM tasks they engage in are intertwined with health – and medication management activities; 4) PHIM activities occur in various locations, at home and away from home, (e.g. physician’s office or while traveling); 5) specific tool design features can be helpful for older adults (e.g. built-in aids to support information comprehension may help overcome age-related cognitive decline).

Discussion. Based on these insights, we aim to propose a research agenda with respect to the themes from the Patient Work Framework and from an overall perspective on the phenomenon.

Title: Longitudinal modeling of prescription refill records to predict medication use

Authors: Kimberley ES Kondratieff*, MS; Candace D McNaughton†, MPH, MD, PhD; Michael E Matheny*, MPH, MS, MD; Thomas A Lasko*, MD, PhD

Description of Work: A majority of Americans now live with chronic diseases requiring long-term management with medication. Successful therapeutic outcomes—reduction of symptoms, reduced morbidity and mortality—require adequate adherence to medication. Similarly, after-market exposure research requires good data on when patients were exposed to medications, including doses received and timing of doses relative to outcomes of interest.

Researchers often use pharmacy refill records as an objective proxy for medication-taking behavior because they are easy to obtain and lack reporting biases. Recent studies[1] suggest analyzing these records as trajectories of adherence instead of single-number summaries improves models predicting patient outcomes[2]. Yet models that predict patient medication adherence trajectories over time remain elusive, requiring months of patient medication records—which may not be available to clinicians—to make accurate predictions[3].

One possible reason for this shortcoming is the dichotomizing of medication trajectories required by current prediction methods. Herein we investigate the use of more granular trajectories to model the association between patient refill behavior and laboratory values over time.

Methodology: Our initial code development project created trajectories for a convenience sample of fifty patients. We then performed in-depth chart reviews for a small number of patients as spot checks on the trajectories’ validity. This included an extended review for one patient where medication lists present in clinical notes were compared to the patient’s calculated supply of medication on the dates of those notes.

For future research, we propose methods for computing granular refill trajectories for two medications (Synthroid and coumadin) and modeling the relationship between these trajectories and longitudinal changes in related laboratory values (TSH/T4 levels and prothrombin time, respectively). We will additionally quantify the level of adherence to these medications—determined from refill records—required to cause clinically significant changes in their associated lab values.

Initial Results: In the in-depth chart review over a single patient, 52% of notes with medication lists neglected to mention a medication the patient had supply for, and 50% of notes included a medication the patient did not have supply for. A spot check comparing another patient’s longitudinal refill pattern of warfarin with INR values over the same time period showed that refill persistence less than 70% tended to correspond with INR values below therapeutic range, suggesting the method’s potential utility.

References


* Department of Biomedical Informatics, Vanderbilt University, Nashville, TN
† Department of Emergency Medicine, Vanderbilt University, Nashville, TN

Maciej Kos, MA, MS1, Christine Gordon, MPH2, Xuan Li, MFA3, Iman Khaghani-Far, Ph.D. 4, Misha Pavel, Ph.D. 5, and Holly B. Jimison, Ph.D., FACMI 6

123456Khoury College of Computer Sciences & Bouvé College of Health Sciences, Northeastern University, Boston, MA, United States

Introduction

Development of just-in-time adaptive health interventions (JITAI) requires reliable technologies for accurate monitoring of physiological parameters in ambulatory settings. While many consumer-grade wearable sensors enable continuous and unobtrusive in-situ heart rate (HR) monitoring, their manufacturers do not sufficiently characterize sensor accuracy. Consequently, researchers independently evaluate HR sensors against electrocardiographs (ECG), using highly heterogeneous and at times inappropriate (1) methods. Thus, harmonization of relevant findings and comparison of sensors evaluated across studies are often limited. To ameliorate this issue, we identified and reviewed suitable methods, characterized their appropriateness for validating HR sensors against ECG, and proposed guidelines.

High veracity of HR estimates is critical for health state estimation, intervention efficacy, and patient safety. Resting HR (RHR) and HR recovery (HRR) are strong, modifiable predictors of cardiovascular health and mortality (2,3). Exercise at 75% of maximum HR helps restore executive function (4): overestimating HR by 10 beats (~5%) can decrease intervention efficacy. This degree of error may put cardiac rehabilitation patients at risk. Small inaccuracies in RHR and HRR estimates may lead to biased health state inferences when comparing patients to clinical benchmarks.

Methods

First, we identified HR sensor validation methods by reviewing literature found in PubMed and Google Scholar by using combinations of the following terms: heart rate, wearables, smartwatches, validation, assessment, accuracy, methods. We excluded articles that focused only on heart rate variability. Second, we applied the methods to synthetic and empirical data collected in an ambulatory setting (5) to rigorously characterize their accuracy and robustness.

Results and Guidelines

HR sensors are validated against ECG using a myriad of methods including various combinations of Pearson's and Spearman's correlation coefficients, t-test, intraclass or concordance correlation coefficients (ICC, CCC), Bland-Altman’s plots, limits of agreement (LoA), and descriptive statistics. Our results indicate that due to its statistical properties, CCC is a preferred method for assessing accuracy. For clinical relevance, LoA should augment CCC. Computing both estimates over a short window may enable using a regression to relate CCC and LOA to motion and HR strata resulting in a better characterization of sensor’s performance. Time lags required for signals’ alignment should be reported as they can jeopardize JITAI’s effectiveness.

Conclusion

The results suggest the importance of CCC, LoA, their regressions on HR strata and motion, and time lags in HR sensor validation studies.

References

Descriptive Comparison of Patient Cohorts with Cognitive Function Testing Results from Structured vs Free-Text Clinical Notes

Aaron Kub1 B.A., Amanda Yoho1 Ph.D.
1 IBM Watson Health Analytics Innovation

Methodology
We used the IBM Explorys Therapeutic Dataset of over 62M patients to identify those with Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) observations. 89,018 patients were identified as having taken any of the two assessments; of these, 16,355 had cognitive assessments from structured EMR fields, while 75,885 had assessments extracted from free-text clinical notes, and 3,204 patients had both types. We compared the age, shared diagnoses, and utilization of patients with any cognitive function testing record from a structured field to those with a record from a clinical note, observing that 3,204 patients were represented in both categories. We further broke down the cohort into those with MMSE structured (MMSE-S) or note (MMSE-N) data and MoCA structured (MoCA-S) or note (MoCA-N) data to compare diagnoses of depressive or neurodegenerative disorders, in order to assess possible impact to studies where cognitive functioning information is important.

Results
Four out of the five most frequent diagnoses for the structured and note cohorts were shared between the two - Depressive disorder, Hyperlipidemia, Hypertension, and Urinary Tract Infection. Uncommon in the two were Anxiety Disorder for the structured cohort and Vitamin D Deficiency for the clinical note cohort. Neurodegenerative diagnoses and depression disorder showed a split between the assessment-specific subgroups, which are outlined in Table 1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>MMSE-S</th>
<th>MMSE-N</th>
<th>MoCA-S</th>
<th>MoCA-N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Neurodegenerative Disease</td>
<td>0.24</td>
<td>0.5</td>
<td>0.44</td>
<td>0.56</td>
</tr>
<tr>
<td>Alzheimer's</td>
<td>0.12</td>
<td>0.25</td>
<td>0.22</td>
<td>0.27</td>
</tr>
<tr>
<td>Parkinson's</td>
<td>0.02</td>
<td>0.08</td>
<td>0.04</td>
<td>0.1</td>
</tr>
<tr>
<td>Depressive Disorder</td>
<td>0.74</td>
<td>0.52</td>
<td>0.53</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Utilization metrics showed a larger deviation between the two cohorts. The average annual rate of visits per patient for the period of 2016-2018 are outlined in Table 2.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Ambulatory</th>
<th>ED</th>
<th>Inpatient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>6.80</td>
<td>0.58</td>
<td>0.63</td>
</tr>
<tr>
<td>Note</td>
<td>10.80</td>
<td>0.16</td>
<td>0.67</td>
</tr>
</tbody>
</table>

We found that 70% of MMSE-S records originated from inpatient psychiatric encounters while 66% of MMSE-N were recorded during office visits. The majority of MoCA scores originated from office visit encounters with 56% and 86% for MoCA-S and MoCA-N, respectively. In total, 50% of all structured records came from inpatient psychiatric encounters and 73% of all note records came from office visits. The note cohort also skewed older, with the average age of patients for structured and note cohorts being 60 and 75, respectively.

Conclusions
In both MMSE and MoCA test types, neurodegenerative disease is more common in patients that have results recorded in clinical notes. Meanwhile, depression disorder is more common in patients with structured results. Suggestions to administer MMSE tests to patients with behavioral emergencies may explain this difference, especially in the context of the increased inpatient psychiatric care utilization observed in the data. Although the patient profiles had considerable differences, more research is required to determine whether these patterns exist in other real-world EHR datasets.

References
An Informatics Infrastructure for Remote Monitoring and Management of Symptoms in Cancer Patients

Gilad J. Kuperman, MD, PhD, Yeneat Ophelia Chiu, MHA, Margarita Rozenshtein, Melissa R. Zablocki, MPH, Claire E. Perry, Abigail Baldwin-Medsker, MSN, RN, OCN, Amandeep K. Dhani, MPH, Brett A. Simon, MD, PhD, Robert Daly, MD, MBA
Memorial Sloan Kettering Cancer Center, New York, NY

Introduction

Patients receiving chemotherapy frequently develop symptoms that, if not managed effectively, require treatment in emergency or inpatient settings. At our institution, 36% of patients started on antineoplastic treatment have an ED visit in 6 months. Previous studies have demonstrated that continuous monitoring of patients’ symptoms can improve outcomes including overall survival. We set about to implement an operational infrastructure to capture patient symptoms on a daily cadence and support the workflows of the dedicated team assigned to monitor and manage these symptoms.

Methods

This work was completed at a single large comprehensive cancer center. Clinical experts developed an interactive questionnaire to capture 15 symptoms from high-risk patients beginning chemotherapy (pain, fatigue, nausea, etc.). The technical components of the infrastructure are: (i) a risk stratification engine that uses ~400 clinical variables to identify patients most at risk for requiring an emergency room visit within 6 months of initiation of therapy, (ii) electronic features that notify the physician that the patient is eligible for -- and allow the physician to enroll the patient in -- the program, (iii) a cohort manager application that keeps track of the patients who should be receiving a daily survey, (iv) a questionnaire driver, housed in our internally developed patient portal, that makes the questionnaire available to the patient, (v) a notification engine that sends electronic messages to the clinical team about concerning symptoms, (vi) a workflow application that allows the clinical team to review symptoms and the list of enrolled patients, and (vii) interfaces to the EHR to create a patient list and other workflow support.

Results

A pilot program went live on October 15, 2018. As of January 23, 2019, 28 patients had been enrolled in the program (average age 65; 36% male). In that time, 1,350 daily questionnaires were sent to patients of which 840 were completed (62%). 548 questionnaire responses included concerning symptoms. The technical platform is performing as intended though minor remediation has been required to assure that the various components interoperate robustly. A dedicated design team is examining user needs to identify opportunities to enhance the workflow support tools.

Conclusion

We were able to architect an environment to deliver questionnaires to patients on a daily basis. The optimal workflows to support capture and management of patient data still are being explored.

Figure. Data flow diagram for remote monitoring of patients started on chemotherapy.
Improving Clinical Decision Support System by Analyzing Prognostic Factors of Colorectal Cancer Stage IV Patients

Hyunjin Kwon¹, Youngho Lee, PhD¹, Jinhyeok Park, BS¹, Sunjin Sym, PhD², Jeong-Heum Baek, PhD²
¹Gachon University, Gyeonggi-do, Republic of Korea; ²Gil Medical Center, Incheon, Republic of Korea

Introduction

Watson for Oncology (WfO) is a medical software created by IBM that combines Big Data and Artificial Intelligence. It is one of the Clinical Decision Support System that provides a suggestion of treatment to the oncologists based on past clinical cases, medical journals, and specialized data. WfO's treatment suggestion shows a high concordance rate with clinicians but shows low concordance rate to metastatic disease. This is because WfO has a critical flaw that does not reflect the EMR data. Thus, this research was done as preliminary research on improving the performance of WfO. To increase the low concordance rate of WfO, not only general guidelines but information based on actual EMR data is needed. To acquire the information from the actual EMR data, we analyzed the prognostic factors that affect the overall survival in patients with stage IV CRC.

Method

In this study, we used the EMR data of Gil Medical Center where first adopted WfO in Korea. We retrieved EMR data of 229 patients with stage IV CRC for analyzing prognostic factors of survival rate. We analyzed the median overall survival using the Kaplan-Meier method. We also identified the prognostic factors which affect the survival rate by using the Cox proportional-hazard models and tested the proportional hazards assumption using Kaplan-Meier curves as well.

Result

The median survival was 1.75 years. The univariate analysis demonstrated that age, number of harvested lymph node, positive rate from harvested lymph node (Ratio), lymphatic invasion, post-operation and chemotherapy (Post OP Chemo) were significant factors affecting survival (p < 0.1). The multivariate Cox regression analysis revealed that age, Ratio, and Post OP Chemo were associated with survival (Table 1).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Hazard Ratio</th>
<th>95% Confidence Intervals</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (year)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 65 vs. ≥ 65</td>
<td>1.81</td>
<td>1.32 - 2.47</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less Than 20% vs. Between 20% and 50%</td>
<td>1.56</td>
<td>1.09 - 2.22</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Less Than 20% vs. Greater Than 50%</td>
<td>3.1</td>
<td>2.05 - 4.69</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Post OP Chemo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present vs. Absent</td>
<td>1.36</td>
<td>0.99 - 1.89</td>
<td>&lt; 0.1</td>
</tr>
</tbody>
</table>

Conclusion

To overcome the limitation that WfO does not reflect the EMR data, the information based on analyzing real EMR data is required. In this study, we have identified the prognostic factors affecting the survival period of patients with stage IV CRC. On multivariate analysis, we demonstrated that age, Ratio, and Post OP Chemo were prognostic factors for overall survival of stage IV CRC. In future studies, we expect that if this result is applied to WfO, it will be possible to offer an improved treatment suggestion in the field of stage IV CRC.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2018-2-00861, Intelligent SW Technology Development for Medical Data Analysis)
Addressing the Medical and Business Challenges in the Creation of the CureCloud™, a Prospective, Linked, Direct-to-Patient Multiple Myeloma Registry

Steven Labkoff, MD, FACP, FACMI, FAMIA, Christopher Williams, Shaadi Mehr, PhD, MS, Eugene Hui, Devon Bush, Scott Longley, Esme O. Baker, Simone Maiwald, David Voccola, Charles Tirrell, Leon Rozenblit, JD, PhD

Multiple Myeloma Research Foundation, Norwalk, CT; Broad Institute of MIT and Harvard, Cambridge, MA; Prometheus Research, New Haven, CT

Introduction: Understanding the onset, progression, and outcomes of patients with multiple myeloma requires linked data from disparate sources. The CureCloud™ is a follow-up project to the Multiple Myeloma Research Foundation’s landmark 1043-person CoMMpass study [1, 2]. The CureCloud Direct-to-Patient Registry (CC-DTP) (NCT03657251) is a prospective, longitudinal, linked data set. The CC-DTP’s goal is to accumulate data on 5000+ multiple myeloma (MM) patients to better understand and characterize this disease and to provide a comprehensive data view for patients, researchers, and clinicians [3]. CureCloud is currently slated to be a $20M effort, approximately $3M of which is earmarked for the informatics infrastructure.

Methods: After IRB approval, workflows were designed to obtain four types of data, patient self-reported data, EHR abstractions (from Cota Healthcare and Tempus), myeloma-specific gene panel sequencing (from the Broad Institute), and exome genomic sequencing (from Tempus) for the initial instantiation. Patients are solicited to join the program in multiple ways: direct outreach via social media, direct-to-patient emails from the MMRF database, and via direct outreach to patients via a partner’s (Tempus) outreach to treating MM physicians and their patients. Developing an overall data architecture became necessary to ensure the complex design with many partners would satisfy all intended use cases. Working with Prometheus Research, a comprehensive work- and dataflow architecture was designed and implemented [3]. This architecture considered major issues such as patient consent, the obtaining and processing of biospecimens for sequencing and biobanking, electronic health record (EHR) data abstraction, and data privacy and stewardship concerns. The final architected program will support large-scale data collection, analysis and reporting.

Results: The overall architecture of the CC-DTP, once designed, was tested with 25 pilot patients. These patients were selected to represent diverse geography in order to test the speed and completeness of the Cota EHR abstraction process. Patients successfully navigated informed consent and biospecimen donation workflows. Pan-cancer genomic and a MM multi-gene panel sequencing were performed; those data were matched with the EHR abstractions and entered into the registry. By late 2019 or early 2020, a planned, CLIA-approved patient report and customized visualizations will be provided back to participants. At the time of this writing, the CC-DTP contains 73 patients and is expected to grow to over 5000 over the next 5 years. Additional matched data sources are planned, including proteomic immune data from existing biospecimens, validated patient-reported outcomes and medical insurance claims. Because the CC-DTP is serving 3 distinct end-user groups, careful segregation of access privileges to preserve patient security and privacy is critical.

Conclusion: The creation of a nationwide, prospective, multi-data type, direct-to-patient registry requires thoughtful architecture. This architecture defines how data will be made available to the right end users with proper attention to patients’ privacy. Researchers will have a comprehensive data set upon which to study disease progression prospectively, to identify new biomarkers, validate therapies, and test new hypotheses to find a cure.

References
MPath Non-NGS: Software for Analysis of Clinical Fragment and Qualitative Polymerase Chain Reaction (PCR) Assays

Sean Lachhander, MS, John Ziegler, MS, Shoham Das, BS, Paulo Salazar, BS, Nana Mensah, MT, MBA, Ivelise Rijo, BS, Lissette Fernandez, MPH, Lin Dong, BS, Daniela Elezovic, BS, Ahmet Zehir*, PhD, Maria Arcila*, MD

1Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY

Abstract
Molecular diagnostic laboratories are a critical part of patient care and are inundated with paper records. Every test, sample, and record increases complexity and makes it more difficult to meet pressing deadlines with accuracy. Furthermore, keeping paper records renders data mining virtually impossible. MPath Non-NGS is a comprehensive set of web-based applications that enables gathering, analysis, reporting and mining of a series of Non-NGS assays performed in MDS at MSKCC.

Introduction
While the next-generation sequencing (NGS) based assays have inherent support from a computational and informatics point of view, the traditional Non-NGS assays have not benefited from this digitization. Here, we describe our implementation of MPath Non-NGS, a set of software systems designed to address this deficiency, at the Molecular Diagnostics Service (MDS) in the Department of Pathology at Memorial Sloan Kettering Cancer Center (MSKCC).

System Overview
MPath Non-NGS software suite targets all stages of clinical workflows, integrating into the laboratory information management system (LIMS), and addresses the needs of lab technologists and supervisors with tools for efficient test setup, result tracking, and supervisor/QC (quality control) review. Subsequent stages of the assays are handled via a molecular result analysis and sign-out application. On these applications, molecular fellows and attending pathologists can view, annotate, and analyze the results of each clinical sample. Some key features include: (1) automated graphing of discrete Short Tandem Repeat (STR) test values and tools for calculating chimerism, (2) one-click human-readable annotation and reporting of high-density single nucleotide polymorphism (SNP) array results based on post commercial software analysis, (3) retrieval and curated views of historical results on a per-patient basis, and (4) automated diagnosis of HER2/IQFISH samples. Once each sample is reviewed, results are stored and indexed for future retrieval and clinical report generation.

Figure 1: Sample result visualization for STR testing
The applications are built utilizing a modern software development stack that includes a non-relational database deployment, python web APIs, and an intuitive web user interface built with Javascript/ES6. MPath Non-NGS is currently deployed on-site at version 2.4 and supports the processing of approximately 750 samples per month across 18 clinical assays. In total, MDS staff has used MPath Non-NGS to process over 26,000 samples in the past three years. Utilizing the software suite has shown to decrease the amount of human errors seen in assay workflows, and has resulted in significant time savings for lab technologists and reviewers. Overall, we are able to report results in an average of two days sooner versus using traditional written records, and have seen a 76% reduction in paper records. Additionally, lab technologists save over an hour per batch of samples. The next steps include supporting additional assays in the MDS and using back-end APIs to share the data across the institute.

Conclusion
In this presentation, we will walk-through typical workflows from the perspective of various users, including molecular pathologists, analysts, lab technologists, and supervisors. We will also provide an understanding of the software development processes and frameworks utilized to build these systems, as well as discuss some of the conclusions drawn while implementing at a large-scale in MSKCC.
Challenges of Data Abstraction to Characterize Contrast Media Reactions and Identify Potential Risk Factors in a Cancer Patient Population.

Cesar Lam, MD1; Christina Eldredge, MD2; James E. Andrews, Ph.D.2; Randa M. Perkins, MD1
1H. Lee Moffitt Cancer Center Research Institute, Tampa, Florida; 2University of South Florida, School of Information, Tampa, Florida

Introduction: Contrast imaging plays a pivotal role in the diagnosis, staging, and treatment of cancer. Cancer patients are frequently imaged to assess response to established complex treatment regimens and, in some cases, to continue in clinical cancer trials. Cancer patients may be at an increased risk for contrast media reactions (CMRs) due to the frequency of exposure to contrast media during these exams1. Additionally, these patients often undergo multiple treatment regimens, which may cause metabolism and immune system changes, potentially placing them at a greater risk for developing CMRs. We sought to use a data-driven approach to determine potential risk factors for these reactions; however, several challenges to automated data abstraction were identified. Furthermore, evaluation of current contrast allergy notes lacked the standardized CMR diagnostic criteria, recommended by the American College of Radiology (ACR)2.

Objective: To identify and address challenges in data abstraction to characterize potential risk factors in patients undergoing imaging at a cancer center.

Methods: The dataset included patients imaged in the radiology department at a cancer center between 2013 and 2017. The quality improvement department registry of CMRs occurring in this time period was used as the basis of the query. The following data points were queried in electronic medical record (EMR) system for the cohort of patients in which CMRs had occurred during the time period: (1) demographics; (2) cancer type and location; (3) treatment; (4) time from first treatment to CT scan reaction date; (5) allergies at time of CT scan reaction date; (6) creatinine at CT scan reaction date; and (7) medications at CT scan reaction date. These factors were collated and descriptively analyzed. To address gaps in data elements, event contrast allergy event notes were also evaluated.

Results: Between 2013 and 2017, a total of 122,997 contrast enhanced CT scans were performed. During this 5 year period, 150 CMRs were documented. The cohort median age was 67 years old with 52% female. Multiple CMRs occurred in 6 patients. The distribution of cancer types were as follows: gastrointestinal 40 (27%); genitourinary 30 (20%); soft tissue 17 (11%); breast 17 (11%); gynecological 8 (5%); endocrine 8 (5%); lymph node 6 (4%); oral 4 (3%); and 2 (1%) were unknown. In 16 reactions (11%) cancer descriptors (data elements) were missing. In the dataset, the abstraction method created duplication of CMR events in patients having multiple cancers (13 patients). Additionally, in 17 patients, no EMR data was identified. Finally, each queried data point was abstracted as a separate Excel spreadsheet making it difficult to analyze without merging it manually.

Conclusion: The key challenges to identifying and characterizing CMRs include lack of a systematic/standardized approach to data collection, duplication of events, and missing data points within the EMR system. Furthermore, the queried output of several pages of data necessitated the formation of a new data spreadsheet manually abstracted, increasing the potential of data entry error.

Future Work: We will analyze the etiology of missing data points and correct systemic flaws. Furthermore, we will evaluate and improve the current data collection and abstraction process using a standardized data abstraction tool with the aim of creating a single comprehensive data set.

References


Towards a learning health system for contraception care: Representing knowledge for tailored clinical practice feedback

Zach Landis-Lewis, PhD, MLIS1, Veena Panicker, BDS1, Dahee Lee, MHI, RN1, Marisa Wetmore, MPP2, Roger Smith, MD2, Michelle Moniz, MD2
1University of Michigan Medical School, Ann Arbor, MI, 2Program on Women's Healthcare Effectiveness Research, Ann Arbor, MI

Introduction
Learning health systems possess advanced knowledge management capabilities that refine knowledge, often in computable forms, through learning cycles. A key step of the learning cycle is the transition from knowledge to practice (K2P), in which refined knowledge is used to generate interventions to influence healthcare professional practice. Clinical practice feedback, delivered via quality dashboards and reports, is a widely used strategy to change practice with a growing evidence base. Using the learning cycle as a conceptual framework, we explore the development of computable knowledge for learning about clinical practice feedback interventions, applied to the implementation of a new practice of offering immediate postpartum long-acting reversible contraception (IPLARC) to patients. Clinical practice feedback about IPLARC includes hospital-level performance measures based on percentages of eligible women who are a) offered contraception, and b) provided with IPLARC at time of childbirth.

We draw from two domains to explore a foundation for computable knowledge in tailored practice feedback: user stories and behavior change interventions. In a prior study we explored the affordances in adopting user stories to manage requirements for the design of 11 tailored performance feedback messages with a collection of 9 user stories for IPLARC. User stories associate a specific software function (what) with a specific user (who) and a rationale (why). Tailored feedback messages could be matched with the recipient’s characteristics, specified in user stories, to deliver appropriate and actionable feedback. Behavior change interventions are “policies, activities, services, or products designed to induce or support people to act differently from how they would have acted otherwise”2. Components of behavior change interventions are specified in the Behavior Change Intervention Ontology (BCIO)2.

Methods
To understand the potential of BCIO for developing and managing computable knowledge about tailored feedback messages, we conducted a preliminary mapping exercise that related IPLARC user stories with BCIO constructs.

Findings
User stories were associated with the following BCIO constructs: Content (n=9); Population (n=7); Engagement (n=2) and Delivery (n=2). For example, the user story "As an attending physician, I want to receive a performance feedback report within the body of an email so that I can spend less time opening attachments" was mapped to Population (Context), Delivery (Intervention) and Engagement (Exposure) constructs. BCIO constructs were useful for organizing requirements and recognizing opportunities to formally specify tailored feedback message knowledge, based on individual and situational difference characteristics captured in the user stories.

Conclusion
User stories and BCIO may be useful as a foundation for computable knowledge representations of tailored performance feedback reports in learning health systems. Future work is required to understand implications for their management at large-scale.

References
Using EHR Data to Evaluate CPOE Safety
Angela L. Laurio, DrPH, RN, Aaron S. Dietz, PhD, Jean M. Scott, MS, CPHIMS, David C. Classen, MD, MSc
1Informatics Patient Safety, U.S. Department of Veterans Affairs, Washington, DC
2University of Utah, Salt Lake City, UT

Abstract
In 2018 the VA piloted the University of Utah EHR Flight Simulator CPOE evaluation tool at three hospitals to assess medication ordering safety and found it insightful, but too time and resource intensive to scale across all 130 hospitals on a recurring basis. This study demonstrates the feasibility of using EHR data to automate and scale CPOE evaluation across the enterprise as a more efficient and sensitive alternative to simulation using fewer resources.

Introduction
Despite the implementation of CPOE, medication errors still occur, often during order entry. The University of Utah EHR Flight Simulator uses simulated order entry to evaluate CPOE performance. The Department of Veterans Affairs (VA) piloted the Simulator at three hospitals in 2018 and found it took 4.5 to 6.5 hours per facility to complete entry of 48 orders. Orders were entered into a test system by one to two providers per site. This study demonstrates the feasibility of using EHR data to automate and scale CPOE evaluation across an enterprise as a more efficient and sensitive alternative.

Methods
The VA Corporate Data Warehouse contains EHR and system setting data. Some settings in the VA EHR are customizable and configuration varies between hospitals. Allergies, inpatient medication orders, and order checks were analyzed for the period 1/1/2017 to 12/31/2018 for the three hospitals that participated in the Simulator pilot. Overall order check counts and reasons were examined. Order check rates were compared for four allergy-drug pairs between simulated order entry and EHR data. Allergy, medication orders, and order check counts were generated for each hospital (Table 1).

Results
Data for 154,438 medication orders and associated order checks from the three hospitals were retrieved in under four minutes. Allergy-drug interaction was the third most common reason for an order check to “fire” (Figure 1). Analysis found a higher order check rate for allergy-drug interaction in EHR data than with simulation at one hospital, but lower order check rates at the other two (Figure 2). Drug classification prevented some order checks from firing, e.g. Sulfa classified as an antibiotic vs. an antacid. Disabled system settings could prevent order checks from firing, but no cases were identified in these data.

Table 1. Order Check Reasons and Counts

<table>
<thead>
<tr>
<th>Allergies, Orders, Checks</th>
<th>Hospital 1</th>
<th>Hospital 2</th>
<th>Hospital 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin Allergy</td>
<td></td>
<td></td>
<td></td>
<td>2,831</td>
</tr>
<tr>
<td>Aspirin or Ibuprofen Order</td>
<td></td>
<td></td>
<td></td>
<td>588</td>
</tr>
<tr>
<td>Order Check</td>
<td></td>
<td></td>
<td></td>
<td>552</td>
</tr>
<tr>
<td>Morphine Allergy</td>
<td>1,071</td>
<td>213</td>
<td>252</td>
<td>1,536</td>
</tr>
<tr>
<td>Oxycodone Order</td>
<td>1,977</td>
<td>130</td>
<td>303</td>
<td>2,410</td>
</tr>
<tr>
<td>Order Check</td>
<td>1,855</td>
<td>130</td>
<td>282</td>
<td>2,267</td>
</tr>
<tr>
<td>Penicillin Allergy</td>
<td>9,452</td>
<td>1,623</td>
<td>4,629</td>
<td>15,705</td>
</tr>
<tr>
<td>Penicillin or Amoxicillin Order</td>
<td>182</td>
<td>49</td>
<td>253</td>
<td>484</td>
</tr>
<tr>
<td>Order Check</td>
<td>136</td>
<td>48</td>
<td>210</td>
<td>394</td>
</tr>
<tr>
<td>Sulfa Allergy</td>
<td>2,866</td>
<td>513</td>
<td>1,150</td>
<td>4,529</td>
</tr>
<tr>
<td>Sucralfate or Carafate Order</td>
<td>18</td>
<td>7</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>Order Check</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Conclusion
Data collection time was reduced by 99% using EHR data versus simulation, while the number of orders analyzed increased by over 3000%. Use of EHR data provides a more realistic assessment of CPOE performance over simulated order entry using test accounts. Further analysis is needed to determine why some order checks didn’t fire and to assess patient outcomes.

References
GeoService: Bridging Clinical and Public Health Informatics

Adam M. Lee, MBA
University of North Carolina, Chapel Hill, NC

Introduction

The North Carolina Translational and Clinical Sciences (NC TraCS) Institute set to establish a service, dubbed “TraCS GeoService”, for researchers at the University of North Carolina - Chapel Hill offering geocoded information along with other clinical data extracted from the UNC Healthcare system Epic electronic health record. This service has three aims 1) Link census data associated with subjects about their census tract from the available variables. 2) Deliver “as the crow flies” distances between a subject’s coordinates and various UNC locations using the Haversine formula. 3) Derive longitudinal data about a subject’s geographical location and census data tracking the number of moves, distance moved, and frequency.

Methods

TraCS GeoService implemented geocoding with the following methods:

1) Reference data are updated and verified. SAS provides street address files for download on a yearly basis, and ZIP codes quarterly. The census demographic data are loaded as of 2014.

2) Addresses are formed and extracted from the clinical data warehouse. Addresses are not cleaned or altered, but invalid ZIP codes are omitted, which remove non- U.S. locations.

3) SAS code executes to process the extracted addresses into batches and runs SAS PROC GEOCODE on each batch.

4) Each batch of addresses are geocoded and imported into the clinical data warehouse as they process, resulting in latitude, longitude, and Federal Information Processing Standards (FIPS) Census Tract codes.

5) Linkages are completed ad hoc for researchers’ data provisions.

Researchers may utilize geocoded data as additional data attributes for data sets, and/or use geocode or census data as inclusion/ exclusion criteria for cohort identification, i.e. recruitment cohorts may require subject to be within 50 miles of a specific facility or target patients in tracts with a certain percentage indicated as ‘Below Poverty Level’

Results

The geocoding process results in street level matches of 74.4% of the population, with three categories for accuracy, ‘Accurate, Minor Error, and Major Error’ matches with matching percentages found in Table 1. The Zip level matches account for 25.5% of the population, and is the general default for non-matches due to poor address formation or lack of information.

Discussion

The primary discussion for the TraCS GeoService is the omission of cleaning addresses. While accuracy could increase with properly collected and formed addresses, it was crucial to the service that data remain as natural as possible as extracted from the clinical system. Notably, these results facilitate the capabilities of each aim and allows researchers to associate clinical data with socioeconomic data. Additionally, results of this process permit further capabilities such as identifying urban/rural areas, and allows tertiary linkages such as environmental data.

Conclusion

Overall, TraCS GeoService established reliable processes for geocoding and meaningful data linkage from census data to add levels of information to patient populations enhancing UNC research informatics pipeline.
How Personalized Are Personalized Order Sets? Measuring Variation in Physician Use of Electronic Order Sets

Jaehoon Lee, PhD1,2; Nathan C. Hulse, PhD1,2

1 Intermountain Healthcare, Salt Lake City, UT; 2 University of Utah, Salt Lake City, UT

Introduction Computerized Provider Order Entry (CPOE) systems and electronic order set libraries are key components in modern EHR systems that support care standardization by reinforcing best practice through templates. However, enterprise standardized order sets often do not cover ad-hoc needs, exceptional cases, and other provider-specific needs. Personalized order sets (‘favorites’) are designed to address these issues and have been implemented in several CPOE systems, affording users flexibility to modify contents from the standard order sets. An important management point in maintaining these libraries is to control personalization within the originally intended use of order sets so as to prevent clinical misuse, avoid potential safety issues, or the proliferation of redundant content. As a baseline study to better address this issue, we studied two research questions: 1) How are our providers tailoring standard and personalized order sets during routine use? and 2) How can we quantitatively measure such variation?

Methods To measure variation from library templates during CPOE usage, we modeled order set use data set as two character strings: 1) the first string consisting of characters for each orderable item in an order set, and 2) the second one consisting of characters for each representing order instances derived from the order set. With this representation strategy, a gap between the two strings is indicative of whether an element of an order set was used as designed or not. To measure the gap, we adopted the concept of Hamming distance, which is a count of differences in positions for which the corresponding symbols are different between two strings of equal length 1. We normalized the measure by dividing by the overall size of the individual order sets as formulated below.

Hamming distance = #orderable items weren’t actually ordered / size of order sets (# of items)

We used data from a CPOE system in our enterprise data warehouse at Intermountain Healthcare. In the CPOE system, providers can order from standard order sets created by clinical departments but can also choose to create and order from personalized order sets that they create themselves to tailor to their needs. We extracted 1,244,950 order sessions (Feb 2015 – Feb 2019) and divided them into two user groups: standard (933,456) and personalized (311,494) in 28 clinical categories. We excluded very small order sets and those ordered less than 100 times. In mapping orderable items to character strings, we only included those are in prechecked status by default.

Figure 1. Distribution of Hamming distances by order sets (colored by clinical category) shows clearly distinguished pattern of standard radiology order sets (ranges 0.5-1), compared to the rest clustered around 0-0.4.

Conclusion Average Hamming distance of personalized use was 0.23 whereas that of standard is 0.306. Our findings are 1) Use of standard radiology order sets exhibited wide levels of variation in use. 2) Excluding that, there was no significant difference in variation during the use of standard and personalized order sets, and 3) The size of order set templates has no strong correlation with overall variation during order set use. Since the aggregation used in this study has limitations, we will conduct deeper analysis at the level of individual order sets, specialties, and providers.

Reference
**Integrated Research Information for patients (IRIS): Creating a Patient Centric Clinical Trial Experience**

Vincent Lee, MBA; Cathy Knowlton, BS; Sai Sindhu Kurra, BS; James White, BS; Katherine Luzuriaga, MD; Jomol Mathew, Ph.D

University of Massachusetts Medical School, Worcester, MA.

**Introduction:** Timely accrual is a challenge for many clinical trials. In a study of 787 phase II/III adult National Clinical Trial Network sponsored trials launched between 2000 and 2011\(^1\), Bennette *et al.* reported that 18% of the studies closed with low accrual or were accruing with less than 50% of target three years or more after initiation. These non-acquiring trials consume resources for setup and operation without contributing to research findings. Lack of awareness and timeliness of information and misconceptions about clinical trials on the part of prospective patients have been attributed to be amongst the root causes of the problem\(^1\). The main public source of clinical trial information (clinicaltrial.gov) contains complex scientific language that is often difficult for the general public to comprehend\(^2\). Additionally, the lack of tools that connect research teams to interested subjects has also remained a significant barrier to clinical trial accrual. ResearchMatch and Trials Today\(^2\) have made significant improvements. A complementary clinical trial portal that can be locally installed at sites with the latest trial information in simplified language would be the key to raising awareness and engaging patients and providers in catalyzing clinical and translational research.

**Methods:** We leveraged the basic principles of the **Attention, Interest, Desire, and Action** (AIDA) model of advertising and marketing to develop a universally adoptable platform – **Integrated Research Information for patients** (IRIS)- for clinical trial sites to advertise trials to patients, connect them to study teams and analyze the site traffic information. Each trial listing can be composed of the **most critical information relevant to patients** such as trial description in layman’s terms, eligibility, time commitment, and required study procedures. Prospective patients can perform **keyword searches** based on a disease group, specific conditions or provider names and retrieve the most relevant trial information. Furthermore, the system **directly connects patients with study teams** by allowing patients to explicitly express their interest in certain trials and provide permission to be contacted by the study teams. The platform’s capability to **integrate with Clinical Trial Management Systems** (CTMS) allows each active trial to be synched up with CTMS so that the most up to date information is available to patients and providers. Finally, patients are encouraged to register and sign up to receive email notifications when new clinical trials open that match their area of interests, thus increasing patients’ ability to discuss trials with their providers when making clinical care decisions if necessary.

**Results & Discussion:** We have currently launched the IRIS platform (conqueringdiseases.org) for all trials at the University of Massachusetts (UMMS). We have established baseline recruitment information to our current trials and will be collecting overall recruitment data from all trials and recruitment rates for selected trials for a year. Data will be used to compare baseline and post-launch enrollment to assess impact. We are currently designing patient and provider surveys and organizing focus groups to obtain feedback from users to further improve the system. Our future development plans for IRIS include volunteer registry, alerting patients who have expressed interest in a disease area about new trials, geographically oriented trial matching, social media based content presentation and EHR integration of IRIS for providers to search for trials that suite their patients.

**References:**

**Acknowledgement:** Research reported in this poster was supported by NCATS of the National Institutes of Health under award number: UL1TR000161 (KL, JM). We gratefully acknowledge Danielle Howard (UMass Center for Clinical and Translational Research), Allison Blodgett (UMMS IRB) for their assistance with operationalization and implementation of the IRIS platform. The content of this publication does not necessarily represent the official views of the National Institutes of Health.
Dialog Act Classification (DCA) for Automated Classification of Online Health Communities

**Young Ji Lee, PhD, RN**, Grace Campbell, PhD, RN, **Victoria Han**, Hyeju Jang, PhD

1University of Pittsburgh, Pittsburgh, PA; 2University of British Columbia, Vancouver, Canada

**Introduction**: Patients and caregivers are using online health communities (OHCs) to seek and follow information from others who have undergone similar experiences. By learning about the topics discussed on OHCs, health care providers can prioritize patients and caregivers’ concerns and are able to develop interventions for the target population. Our long-term goal is to design an automated system to classify needs of patients and/or caregivers written in the OHCs; however, there is no framework for guiding system development. As a first step, we applied the discourse dialog act classification (DCA) to develop the framework for the system.

**Method**: Data comes from the OHC of ovarian cancer sponsored by the American Cancer Society Cancer Survivors Network (http://csn.cancer.org). The forum contains approximately 55,000 posts written by 2,000 users between 2006 and 2016. We selected initial posts of all users, then excluded the ones that were replies to other posts. The process resulted in a dataset of 856 posts in total. We manually identified whether a post was written by a patient or a caregiver. Then, these posts were classified by 12 needs based on the literature (Physical, Psychological/emotional, Family, Social, Interpersonal, Practical, Daily living, Spiritual, Health information, Patient – clinician communication, Cognitive, Miscellaneous), and only included postings with the most common need (health information need; n=447). Each post was broken down the unit of sentences. Two annotators (YL, VH) annotated each sentence with a dialog act classification, Requesting Information, Explaining Information, Expressing an Opinion, Describing an Experience, Providing a Suggestion. Inter-rater agreement was measured between the two annotators using F1 score, a measure of a test’s accuracy.

**Result**: Among the 447, 229 were written by patients and 179 were written by caregivers. Inter-rater agreement of the classification was 0.51. Most of the disagreement came from “explain information” and “describing an experience”. “Request” was most agreed upon. Main difference between the patient posts and the caregiver posts was the sub classification of describing an experience. Patients mostly discussed their own experience (1,213 vs 42) while caregivers discussed another’s experience (808 vs 45) compared to their own.

| Table 1. Result of Dialog Act Classification on Online Health Community among Ovarian Cancer |
|----------------------------------|--|------------------|
| Category                        | # of sentences (patients) | # of sentences (caregivers) |
| Requesting Information           | 473                        | 283                      |
| [1] why this is happening        | 268                        | 172                      |
| [2] what to do                   | 205                        | 114                      |
| Explaining Information           | 1006                       | 401                      |
| Expressing an Opinion            | 107                        | 45                       |
| Describing an Experience         | 1,255                      | 45                       |
| [1] self-experience              | 1,213                      | 808                      |
| [2] someone else’s experience    | 42                         | 853                      |
| Expressing Emotion               | 114                        | 44                       |
| [1] what happened/is happening   | 86                         | 4                        |
| [2] what is going to happen      | 29                         | 48                       |
| Suggestion                       | 2                          | 0                        |
| Other                            | 184                        | 117                      |

**Conclusion**: Only a moderate agreement between two annotators suggests that dialog acts must be clarified before moving forward, particularly “describe” and “explain”. Difference in subcategory between patient and caregiver posts suggests separation between the two is necessary for accurate prediction in the computational model. “Suggestion” category did not use very few times and it will be a category extremely hard to predict when using automated classification. This is the first step of the algorithm development, and DCA will guide us what to consider for the algorithm development. We will improve our classification framework for the future algorithm.

**References**

A User-Centered Model for Electronic Patient-Reported Outcomes Visualizations in Clinical Practice
Cynthia LeRouge, PhD¹², Jenney Lee, MA¹, Courtney Segal, BA¹, Elizabeth Austin, MPH¹, Danielle Lavallee, PharmD, PhD,¹
¹University of Washington, Seattle, WA; ²Florida International University, Miami, FL

Introduction: Clinical integration of Patient Reported Outcomes (PROs) augments routine healthcare, quality improvement initiatives, and population health management with direct insight into the patient experience. As use of PROs increases across healthcare settings, there is a need to develop electronic PRO (ePRO) tools that provide patients the means to answer PRO surveys electronically, efficiently integrate data into clinical workflows, and incorporate effective mechanisms (e.g., visualizations) for reporting PROs to providers. Indeed, many providers perceive ePROs as a potentially valuable addition to augment their patient care processes if it meets preferences for reporting visualizations and if ePRO workflow is streamlined. Robust literature exists to guide the design of visualizations in other domains, but available models do not adequately address considerations specific to the unique nature of ePRO reporting in clinical practice. The objective of this qualitative study is to develop a stakeholder-informed model of critical design considerations for ePROs visualizations that support successful clinical integration of PRO data. This model will assist both ePRO developers and implementation teams in key decisions regarding the content, design, and functions associated with ePROs across a healthcare system.

Methods: As part of a larger project on health system governance, implementation, and reporting of ePROs, our multidisciplinary team performed a review of published literature on reporting and visualizations for ePROs in clinical settings (n=82 manuscripts). We then performed key informant interviews with 20 providers who had experience using a wide range of PROs, using a semi-structured interview guide to explore providers’ design needs and preferences for ePRO reporting. Interviewees spanned 11 medical and surgical specialties at a large academic medical center, making this endeavor the broadest cross-disciplinary examination of PROs use to our knowledge. Interviews were recorded and transcribed. The literature was coded inductively (open coding) for insights about ePRO reporting. Emergent themes pertained to report content, functionality, and visualization design. This analysis formed an a priori coding schema for the interview data, with codes added as additional key concepts emerged. Purposive, axial analysis of the combined literature review and interview code sets resulted in themes supporting the components of our ePRO reporting design model. Constant comparison methodology and 100% code review (by a non-coder) was used to establish rigor and enhance validity for all qualitative analysis.

Findings: Analysis of the literature and key informant interviews resulted in 12 key themes comprising three critical domains related to 1) data and information (i.e., what type of data is provided), 2) system function and interactions (i.e., which capabilities are provided), and 3) presentation (i.e., how data is visually represented), for ePRO tools. Our integrated framework is summarized in the figure below. Details associated with each of these themes highlights key commonalities where standardization is possible as well as situational nuances where flexibility may be needed.

Conclusion: Our findings reveal 12 core areas that providers consider critical to the usefulness and usability of ePRO reporting tools. The integrated framework we propose represents an important contribution to the PRO literature by highlighting common targets for ePRO tools as well as key points of customization needed to support various specialties and provider preferences. Findings also help to position ePRO visualization design among human computer interaction literature related to health informatics. This work examines providers as stakeholders in designing ePRO reporting tools because as “front line” users they are key to successful clinical integration of ePROs. Future work should address the needs and preferences of additional stakeholder groups (including patients). Future work should also consider the applicability of these findings across other medical and surgical specialties.

Figure 1: An Integrated Framework for the Design of ePROs Reporting Tools

Figure 1: An Integrated Framework for the Design of ePROs Reporting Tools

1653
Data Requirements for a Patient-Centered Learning Health System

Lysanne Lessard, PhD1, Antoine Sauré, PhD1, Agnes Grudniewicz, PhD1, William Gardner, PhD, MS, AB1,2, James King, MSc, MD, FRCP1,2, Wojtek Michalowski, PhD1, David Murphy2, Kathleen Pajer, MD1,2, Elyse Schipper3, Sinthuja Suntharalingam, MD, FRCP1,2, Raphaël Ménard-Grenier1, Michael Cheng, MD, FRCP 1,2
1University of Ottawa, Ottawa, Ontario, Canada; 2CHEO, Ottawa, Ontario, Canada; 3Parent’s Lifeline of Eastern Ontario, Ottawa, Ontario, Canada

Introduction
This research is part of an ongoing effort aiming at transforming healthcare organizations into learning health systems (LHSs)1. LHSs create organizational and technological structures that can leverage the vast amount of patient data stored in electronic health records (EHRs) to improve the safety, efficiency, and effectiveness of traditional health systems. Despite the widespread acknowledgement that LHSs should ultimately improve patient outcomes, current models of LHSs are not truly patient-centered2. Existing EHRs do not support the collection of data related to the preferences and goals of individual patients, nor do they provide the means for measuring whether they are achieved. Child and youth mental health (CYMH), which poses a massive burden on society, is a prime target to benefit from patient-centered LHSs; unfortunately, current processes and tools in CYMH care do not enable continuous and rapid learning leading to patient-relevant outcomes.

Data Requirements for a Patient-Centered Learning Health System
This project aims at identifying the data requirements and characterizing the data lifecycle for a patient-centered LHS for CYMH. It will focus on learning about which treatments are most likely to have positive impacts from the patient’s perspective. We propose to use Design Thinking, an approach that provides methods to engage patients’ families and care providers as equal partners in the study1. This project will be organized in three phases:

1. Understand stakeholders’ challenges and perspectives. This phase aims at developing a deep understanding of obstacles faced by patients’ families and CYMH care providers in relation to choosing among treatment options and learning from treatment outcomes. Design thinking sessions comprised of individual interviews and activities such as the creation of journey maps and prototype decision-support tools to overcome perceived obstacles will be conducted with these stakeholders. Data collected during design thinking sessions will be analyzed through thematic analysis.

2. Ideate and prototype potential solutions. This phase aims at developing conceptual and practical solutions for implementing a patient-centered LHS for CYMH. We will use the results from Phase 1 to articulate data requirements and a data lifecycle that can leverage the capabilities offered by advanced analytics in order to move from patient-identified goals to rapid and continuous learning that improves patient-relevant outcomes. Process mapping and conceptual data model design will be used to identify the data requirements for this learning cycle. Identified requirements will be used to create a representative data set of cases of CYMH patients amenable to predictive and simulation modeling.

3. Iteratively improve solutions based on stakeholder feedback. This phase aims at arriving at validated data requirements. A set of design thinking sessions following principles of user studies will be held to gather feedback from stakeholders. Participants will be asked to interact with treatment predictions and learning cycle mock-ups in order to assess if they meet their needs and expectations regarding a patient-centered LHS. The mock-ups and underlying requirements will be revised until stakeholders assess them as being adequate for their needs.

Conclusion
The results of our project could be integrated into EHRs to ensure that they support the achievement of patient goals alongside clinical ones. They thus stand to improve health outcomes and patient satisfaction in CYMH care without imposing additional strain on the limited resources available to face the growing needs of this population.

References
Finding the evidence base using citation networks: Do 300 to 400 U.S. physicians die by suicide annually?

Tiffany I. Leung, MD, MPH, FACP, FAMIA, Sima Pendharkar, MD, MPH, FACP, Chwen-Yuen Angie Chen, MD, FACP, FASAM, Michel Dumontier, PhD

1 Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands
2 The Brooklyn Hospital Center, Icahn School of Medicine Mt. Sinai, Brooklyn, NY, USA
3 Department of Primary Care and Population Health, Stanford University, Stanford, CA, USA
4 Institute of Data Science, Maastricht University, Maastricht, The Netherlands

Introduction

An eye-catching estimate declared and parroted in scholarly work and popular media is that “300 to 400 U.S. physicians die by suicide annually.” Reported physician suicide incidence varies depending on method of ascertainment, country, specialty, stage of profession, and gender. However, the data and methods on which this claim is based have not been clearly cited. Valid citation of scientific evidence in an ecosystem of enormous volumes of literature is important to mitigate the propagation of inaccurate claims. Citation distortion occurs when a claim is stated as fact yet actually has no supportive empiric evidence. How can we keep track of the provenance and accuracy of claims regarding physician suicide? One approach is to construct a citation network, representing claims made in publications, their supporting references, data, and other components. By tracing the lineage of a single scientific claim, citation distortion can be demonstrated with successive publications. We constructed a claim-specific citation network focusing on one claim about physician suicide deaths. This is a proof-of-concept to capture evidence that can be used to track estimated rates of physician suicide.

Methods

Claims were extracted from a corpus of articles obtained from a scoping review of peer-reviewed literature on physician suicide. The review had included a variety of article types published between 1903 and 2018 from 37 countries. We manually curated from each article the claim (about the number of physicians who die by suicide annually), claim attribute (author), cited reference(s), and cited data or methods, if provided. Each set represents a type of micropublication, which is a semantic model for scientific claims and evidence that enables knowledge discovery and inference across networks of information. We constructed each micropublication using extracted data, and then combined them into one citation network representing the claims, their relationships, and supporting references and data.

Results

Twenty-three articles, including 14 opinion articles, were identified; 11 cited zero references in relation to their claim about the number of physician suicide deaths annually. Four articles claim a different number of annual physician suicide deaths than the claim of interest. Tracing the claim’s lineage back in time in a manually constructed citation network, based on Greenberg, reveals inadequately supportive data or methods.

Discussion

The medical community and advocates need accurate and trustworthy estimates to communicate the urgency of addressing physician suicide. This study is remarkable because: (1) citation networks have never been applied to suicidology, especially among physicians, and (2) this analysis suggests that the oft-quoted claim that “300 to 400 U.S. physicians die by suicide annually” does not have adequately cited supporting data or methods. These findings suggest that inadequately supported claims about physician suicide can easily propagate through publications. Also, advocates and researchers in this area should be vigilant of potential information distortion. A key limitation is that we traced a single verbatim claim, which in future work, will include additional incidence data and calculated estimates. Our approach of building citation networks about physician suicide appears promising for future studies and for ensuring preserved accuracy of scientific work in this field.

References

Ethical Considerations for Social Media Research
Deborah Lewis, EdD, CRNP, MSN, MPH, Walden University, Minneapolis, MN
Patricia Salai RN, MSN, CNN, CRNP(retired), VA Medical Center, Pittsburgh, PA

Overview
The adage that a “picture is worth a thousand words” illustrates that a single photograph can convey complex meaning. This concept easily translates to visual data produced by social media platforms such as Instagram. Instagram provides a unique visual window into individuals’ lives and can provide insights into the meanings represented by the visual images posted.

As part of an ongoing research project, Instagram images were collected to identify the meaning of those images for rheumatoid arthritis, a chronic disease. Because the information is considered public and did not require any communication with the individual who posted it, by definition this project did not involve human subjects and was approved, as such, by the University IRB. This poster represents one aspect of our ongoing research that raised questions related to the methodology and ethical nature of social media findings.

Problem Identified
Anyone 13 years of age and older can create an Instagram account to share photos and video. The user chooses their user name and decides the level of privacy associated with their account. Collecting Instagram images, for research, using the public search tool, should not result in access to any information marked as private by the individual who posted. Users may choose to post their personal user information publicly, which could be collected via a public search.

For our current study, Instagram images were collected, coded, and analyzed independently by two coders. The initial intent of the research was to identify the meaning represented by the images. As the research progressed, the coders encountered an element of ethical discord, and while adhering to all current ethical standards, became uncomfortable with the idea of sharing human images as part of the results. We ultimately chose not to present images that included human faces. We reviewed the literature and found that in other published research using Instagram images, the images were altered to obscure the identity of human faces to protect the privacy of the persons who posted.

Questions for Future Research
Visual ethnography, our methodology, provided an opportunity to examine the meaning of the visuals utilizing the entire image. While providing rich information, this methodology also raised some unanticipated questions. Was it appropriate to share the full images? Would sharing the complete image, including the facial expression, strengthen the conclusions? If the research consumer only sees blurred or blocked images or no images at all, how can they fully understand the findings of a study based on visual data? Even though social media platforms are in the public domain, and include privacy notifications, are individuals who post images fully aware that their images may be used in research? How is the collection of social media images for research different than photographs taken in public settings? If Instagram images are personal art, are they protected by copyright?

In any qualitative approach, the trustworthiness of findings can also be a challenge. Using public social media searches makes it difficult to identify the origin of the images; therefore, it is difficult to member check with the individual who posted the image. If you don’t member check with the individuals who posted are you accurately reflecting the meaning represented by the image shared? Are the images truly posted by those who have the disease? Ongoing research is important in understanding how social media may contribute to health information sharing and the impact on health outcomes. Future research needs to examine the best approaches for ensuring ethical research methodologies and the best ways to share visual data that results from social media healthcare research.

References
Increasing Transparent and Accountable Use of Data by 
Quantifying the Actual Privacy Risk in Interactive Record Linkage 
Qinbo Li\textsuperscript{1} Adam G. D’Souza\textsuperscript{2} Mahin Ramezani\textsuperscript{1} Cason Schmit\textsuperscript{1} Hye-Chung Kum\textsuperscript{1} 
\textsuperscript{1}Texas A&M University, College Station, TX \textsuperscript{2}University of Calgary, Calgary, AB, Canada

Introduction

In biomedical research, there is a constant tension between the need to leverage data to promote social good and the legal requirements to protect patient privacy by guarding against unnecessary disclosure and use. Record linkage, or patient matching, poses a particular difficulty for information privacy. The goal of record linkage is to integrate heterogeneous databases lacking a common identifier, in order to reliably combine patient data from multiple sources.

Recently, there has been increasing interest in using encryption-based privacy-preserving record linkage methods when there are no better alternatives. However, these methods are primarily designed to compute a predetermined linkage function in a secure manner. In reality, in many projects the linkage function cannot be predetermined without interacting with the data being integrated, making it difficult to have good match rates. In fact, in most record linkage projects it is very important to have someone interacting with the data to clean and standardize the data, build training datasets, tune parameters, and resolve ambiguous links. Typically, the amount of data being disclosed during this interactive linkage process is only a small fraction of the data being linked.

Objective

In this poster, we investigate how to best monitor and quantify the actual privacy risk that occurs during this interaction with the data. We first demonstrate the desired properties of a privacy risk score for interactive record linkage that will allow project managers to better communicate and decide on the best tradeoff points between information disclosure and accuracy of record linkage for a given project. We then propose a measure, k-Anonymized Privacy Risk (KAPR) score, and prove that it meets all the required properties. We further illustrate the utility of KAPR via an example and its use in a prototype privacy enhanced interactive record linkage system called MINDFIRL (MInimum Necessary Disclosure For Interactive Record Linkage).

Methods

There are several desirable properties of the measure for quantifying the privacy risk in interactive record linkage. First, it must measure the actual identity disclosure risk based on only the data that has been disclosed to someone during the interactive record linkage process. Second, the measure should range from 0 (when no data is disclosed) to 1 (when all data has been disclosed and identified). Third, it should be a norm (a non-negative, homogeneous function of a suitable representation of the information disclosure state that obeys the triangle inequality) so that it can be assigned the geometrical interpretation of the “length” of the information disclosed, where “longer” means “higher risk of identity disclosure”. The norm property guarantees that the measure is only dependent on the information disclosed by the user, and not on the precise order of disclosure.

Results

We propose the k-Anonymized Privacy Risk (KAPR) score, which is based on the anonymity set size and the total information disclosed. The anonymity set size is the total number of records that share the same information that has been disclosed. For example, if there are 10 patients named Bob, than disclosing Bob would result in the anonymity set size being 10. The KAPR score is formally defined as $K(\kappa; X) := \frac{\kappa}{ND} ||X||_{1,1} = \frac{\kappa}{ND} \sum_{i=0}^{2n-1} \frac{1}{k_i} \sum_{j=0}^{D-1} |p_{ij}|$, where $X$ is the suitably-defined information disclosed from a total of 2n records with D attributes, $p_{ij}$ is the proportion of characters of attribute $j$ of record $i$ disclosed, $\kappa$ is the minimum allowed anonymity set size and $k_i$ is the anonymity set size of record $i$ based on the current information disclosed.

Conclusion

KAPR score is one of the main components of the MINDFIRL system design. To the best of our knowledge, MINDFIRL is the first system designed to meet the legal requirements for minimum necessary disclosure by disclosing only the needed information for high quality record linkage using an interactive on-demand incremental interface and then to quantify the actual privacy risk based on what PII was disclosed. In addition, it supports setting a pre-specified limit on the total level of disclosure by using the agreed on KAPR budget. We have conducted multiple user studies that have demonstrated that use of MINDFIRL reduced privacy risk to only KAPR=7.85%, compared to 100% when all data is disclosed, with little to no impact on decision quality or completion time.
A Bibliometric Analysis of Research Publications that Adopted the Medical Expenditure Panel Survey (MEPS) Data

Jiacheng Liu, BS, Fei Yu, PhD, MPS
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Abstract

The Medical Expenditure Panel Survey (MEPS) is a large publicly available information source provided by the Agency of Healthcare Research and Quality (AHRQ). We systematically retrieved, screened, and conducted a bibliometric analysis for the research publications that adopted MEPS data. Through a bibliometric approach, this study meant to provide insights to inform policy makers, grant agencies, and researchers of the landscape through the research effort that MEPS data have supported.

Introduction

Since 1996, the Agency of Healthcare Research and Quality (AHRQ) has conducted Medical Expenditure Panel Survey (MEPS) by taking samples of households throughout the United States. MEPS data provide an unparalleled degree of details about the use and payment of health care services. Although more than one thousand publications have used MEPS data in the past two decades, it still shows the lack of an overview on the research effort that MEPS data have supported. This study aimed to systematically survey the research publications that adopted MEPS as the sole or one of the datasets by a bibliometric approach. The findings meant to provide insights to policy makers, grant agencies, and researchers in terms of data collection and organization, funding, publishing, and collaborations.

Methods

This study systematically searched peer-reviewed articles from 1996 to 2018 which adopted the MEPS data in two research databases (PubMed & Web of Science (WoS)). JL and FY independently screened the title and abstract fields of each retrieved publication in Covidence\(^1\) based on a set of inclusion and exclusion criteria. The full citation records of the included publications were obtained by matching records in WoS using PMID, DOI, and title fields. WoS analytics, Microsoft Excel, and VoSviewer\(^2\) were utilized for quantitative and bibliometric network analysis, including the number of publications by year, research categories and topic clustering, journal sources, other datasets co-used with MEPS, funding sources, and co-authorship (i.e., organizational and country collaboration).

Results

A total of 1670 publications from 1996 to 2018 were included in this study. The publications that adopted MEPS data grew steadily with an average of 75 articles per year and reached a peak in 2017. The top research categories were healthcare services, public environmental occupational health, health policy, and internal medicine. In addition, 235 key terms were extracted from the title and abstracts of the included publications and formed three major topic clusters (i.e., health quality and cost, healthcare accessibility and disparity, child and family care). Additional 24 datasets were identified that were frequently co-used with MEPS by the included studies such as Nationwide Inpatient Sample and Truven MarketScan. National Institute of Health was the top grant agency for the included studies. More than 3200 authors contributed to the included publications, and they were from 25 countries and 1074 institutions or organizations.

Discussion & Conclusion

Although MEPS focuses on the representative samples of domestic households and researchers in the U.S. contributed the most to the included publications, the MEPS data have inspired cross-country collaborations. Our results illustrated the broad landscape of research efforts that MEPS data have supported and substantiated the value of AHRQ’s effort of providing the MEPS data to the public. The findings will help stakeholders with informed decision making.

References

Clinical Informatics Course: Doctor of Medicine Student Perceptions

Jialin Liu, MD1,2*, Siru Liu PhD Candidate3, Wenyao Annie Wu4, Tao Zheng BS1
1Department of Medical Informatics, West China Medical School, Sichuan University, Chengdu, China
2Department of Otolaryngology, West China Hospital, Sichuan University, Chengdu, China;
3Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA;
4Department of Biomedical Engineering, Tufts University, Medford, Massachusetts, USA

Abstract

This paper aims to investigate doctor of medicine students’ perceptions regarding a clinical informatics course at West China School of Medicine, Sichuan University. Students were surveyed to evaluate their perceptions and satisfaction of the course. Semi-structured and open-ended questionnaires were used to explore the students’ perception of the course. A total of 95% (195/205) completed questionnaires were returned on October 21-23, 2018. The results showed that students were satisfied with and received benefits from the course.

Introduction

With the expanding interest and investment in health information technology by hospitals, health systems, and practitioners, physicians will be practicing in a digital world. Therefore, clinical informatics course will be of paramount importance for physicians [1,2]. It is important for students in medicine and health care to be taught clinical informatics. This pilot study collected and analyzed quantitative and qualitative data to evaluate doctor of medicine students’ perceptions about a clinical informatics course at West China School of Medicine, Sichuan University. West China School of Medicine is one of the most famous medical schools in China.

Methods

We developed a 9-question with close-and open-ended questions, web-based, anonymous survey in consultation with medical educators and informatics experts. Completion of the questionnaires was voluntary and anonymous. We surveyed all students (205) joining the course on October 21-23, 2018.

Results

The survey response rate was 95.12% (195/205), of which 154 (75.12%) answered the open-ended question. 52.68% of the respondents were male. 94.36% (184) were very satisfied while 4.62% (9) were satisfied. The mean overall satisfactory was 4.92 ± 0.35 (5-strongly agree to 1-strongly disagree); “the course benefits your career” was 4.92 ± 0.39. 94.87% (185) strongly agreed and 3.59% (7) agreed (Table 1). The main responses to the open-ended questions were “clinical informatics is very important”, “I hope to increase the opportunity of clinical informatics learning”.

Table 1. Student perception about clinical informatics course

<table>
<thead>
<tr>
<th>Question</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall satisfaction</td>
<td>184 (94.36%)</td>
<td>9 (4.62%)</td>
<td>1 (0.51%)</td>
<td>0 (0%)</td>
<td>1 (0.51%)</td>
</tr>
<tr>
<td>Benefit your career</td>
<td>185 (94.87%)</td>
<td>7 (3.59%)</td>
<td>2 (1.03%)</td>
<td>0 (0%)</td>
<td>1 (0.51%)</td>
</tr>
</tbody>
</table>

Conclusion

Findings gleaned from this study suggests that 98.98% (193/195) of the students were satisfied with the clinical informatics course. 98.46% (192/195) of students agree that the course will benefit their career. Clinical informatics is a very useful subject for the doctor of medicine students.

Funding: This project was supported by the National Natural Science Foundation of China (Grant Number:71273182)

References


Learning Objective: Understand the doctor of medicine students’ perceptions about a clinical informatics course.
One Size Does Not Fit All – Lessons Learned from Developing Algorithms for Decision Support for Symptom Management in Cancer Patients

David F. Lobach, MD, PhD, MS1; Mary E. Cooley, PhD2; Barbara Halpenny, MA2; Jennifer Ligibel, MD2; Andrea Enzinger, MD2; and Janet L. Abraham, MD2

1Klesis Healthcare, Durham, NC; 2Dana-Farber Cancer Institute, Boston, MA

Introduction
Cancer survivors represent 5.0% of the United States population, and their numbers are projected to increase by 29% over the next 10 years to more than 21 million individuals.1 Most cancer patients are afflicted with distressing symptoms that are undertreated and not managed in accordance with evidence-based practices leading to increased morbidity, mortality and cost.2 Clinical decision support (CDS) can facilitate symptom management in cancer patients; however, studies of CDS have been limited in their scope and scalability.3,4 This study is part of a larger project funded by the National Cancer Institute (NCI) to promote care continuity across clinical settings through CDS for evidence-based cancer symptom management. Development of algorithms for fatigue and constipation were specifically requested in the NCI contract. We engaged clinicians from multiple disciplines who manage patients with cancer symptoms in a variety of contexts to develop generalized algorithms for the management of constipation and fatigue across the continuum of care. This poster reports on feedback from the clinical experts who created the algorithms and from multidisciplinary clinicians who reviewed the completed algorithms.

Methods
We leveraged our previous approach for adapting research evidence and guidelines into clinical algorithms for cancer symptom management using the ADAPTE process specifically targeting fatigue and constipation.5 To receive input from the continuum of clinical settings in which cancer care is provided (i.e. from primary care through acute cancer treatment and end-of-life care), we formed two panels of eight clinicians and other stakeholders, one to address constipation and the other to address fatigue. Panel members included medical oncologists, oncology nurse practitioners, palliative care physicians, integrative care specialists, psychologists, social workers, pharmacists, primary care internists, gastroenterologists, nutritionists, and patients who are cancer survivors. Panel experts were provided current research evidence and guidelines. They interacted in three virtual sessions and one in-person session to review draft clinical algorithms. Themes and observations resulting from these sessions were recorded by study team personnel and used to modify the algorithms. Each modified algorithm was then reviewed with 8 to 10 multidisciplinary clinicians (MDs, PAs, NPs) through structured interviews.

Results
Salient observations from sessions with clinical experts and interviews with multidisciplinary clinicians included:
1) Cancer symptom management guidelines from different reputable sources are conflicting, inconsistent and need more specificity in order to be used for clinical decision support.
2) Recommendations need to be based on a patient’s place on the cancer care continuum, care setting, and patient preferences/goals of care, treatment status/regimen, age, and cancer type.
3) Primary care clinicians desire a general symptom management guideline that is applicable to all patients.
4) Patients of hospice clinicians may have different goals from those of actively treated or advanced cancer patients.

Discussion
Study observations have implications for cancer care delivery models that extend across the care continuum, especially if fewer patients follow-up through oncology clinics as the number of cancer survivors increases. Implications are that CDS for patients being treated for cancer may not generalize to primary care providers, and thus primary care providers should be explicitly included in usability testing. Implications for managing cancer patients in hospice programs are that guidance needs to be aligned with overall care goals and tailored to the changing status of patients as they approach end-of-life. This study is limited in that it reflects only the opinions of 16 clinical experts and 17 multidisciplinary clinicians, and that it focused only on fatigue and constipation.

Conclusions
While CDS can be helpful with promoting evidence-based cancer symptom management, the CDS systems providing care recommendations will need to be specifically developed to fit the patient characteristics and context.

Acknowledgements: This project was funded by the National Cancer Institute contract 75N91018C00022/HHSN261201800022C.

References
Classification Types: A New Feature in the SPECIALIST Lexicon

Chris J. Lu, PhD1,2, Amanda Payne, PhD1,2 and Dina Demner-Fushman, MD, PhD1
1National Library of Medicine, Bethesda, MD 2Medical Science & Computing, LLC, Rockville, MD

Introduction

The SPECIALIST Lexicon (thereafter, the Lexicon) [1], distributed by the National Library of Medicine (NLM) as one of the Unified Medical Language System (UMLS) knowledge sources, supports popular NLP tools, such as SemRep, MetaMap, cTAKES, CSpell, and the SPECIALIST Lexical Tools, as an underlying resource. A new enhanced feature called the classification type (CT) is a proposed addition to the Lexicon. These classification types can be archaic, source, informal, or other. First, terms classified as archaic, such as cozen, colde and benight, are considered no longer in common use in modern corpora (such as MEDLINE). These terms may have modern equivalents in the same lexical record (colde for cold) or in separate ones (ye for the). Second, normalization on spelling variants from foreign English into US English is needed if the source is from a foreign country. For example, British English (analyse, leukaemia, tumour) can be normalized to US English (analyze, leukemia, tumor). These terms are classified as source. Third, consumers often use informal language when they ask questions. For example, bomb for success, or grandpa for grandfather are used primarily in colloquial contexts. The performance of automated consumer question understanding could be improved if the Lexicon provides informal terms with their cross-referenced (CR) formal terms (synonyms). Four CTs and their syntax are shown in Table 1.

Table 1. Enhanced Features of Four Classification Types.

<table>
<thead>
<tr>
<th>Code</th>
<th>Examples/Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>class_type=archaic</td>
<td>base form</td>
</tr>
<tr>
<td>class_type=ar</td>
<td>ched</td>
</tr>
<tr>
<td>class_type=source</td>
<td>base form</td>
</tr>
<tr>
<td>class_type=source</td>
<td>analyse</td>
</tr>
<tr>
<td>class_type=informal</td>
<td>base form</td>
</tr>
<tr>
<td>class_type=informal</td>
<td>bomb</td>
</tr>
<tr>
<td>class_type=other</td>
<td>other type of classification (i.e. gene, protein, etc.)</td>
</tr>
</tbody>
</table>

Classification Type Implementation

CTs are added to new lexical records during lexicon building by NLM linguists through a Web-based tool, LexBuild. New GUI components for adding CTs and the enhanced LexCheck software for validating syntax and contents of CTs are integrated in LexBuild. CT tagging on existing lexical records is also done through LexBuild. First, class_type=unassigned is added to all existing records, and removed after it is tagged. Computer-aided features for retrieving records by specified patterns, like suffix, substring, category, etc. are implemented for systematic tagging. Finally, post-process programs are implemented to generate new Lexicon tables, including archaic terms, spelling variants with originated sources and informal terms with their formal synonyms, for the Lexicon annual release.

Applications and Conclusion

The performance of NLP applications that use the Lexicon is expected to be improved with the new CTs. First, archaic terms can be excluded when dealing with modern corpora. Second, foreign English can be normalized to US English with source CTs. Third, the performance of automated question understanding on consumer health can be improved. For example, grandpa (no concept found in UMLS) can be effectively mapped to grandfather (C0337475) with query expansion by substituting formal synonyms for their informal terms [2]. The Lexicon is distributed by NLM via an Open Source License agreement and is available at: https://umlslex.nlm.nih.gov/lexicon.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. We thank Dr. L. McCreedy, D. Tormey and A.C. Browne for their valuable discussions.

References

Dissemination of HIV-HCV-STD Clinical Evidence through Guideline-Driven Interactive Case Simulation Tools – An Assessment of User Engagement

David Lukacs, MS; Meredith Abrams, MPH, MS; Dongwen Wang, PhD
Arizona State University, Scottsdale, AZ

Abstract

Engaging clinician users in educational initiatives is critical for effective knowledge dissemination. We analyzed the usage of five interactive case simulation tools in a statewide HIV-HCV-STD clinical education program to assess user engagement. We have found that interactive, patient-specific simulation can engage clinician users more effectively when compared to hyperlinked text or static cases. The results from this study will guide the future development of online resources for more effective dissemination of clinical evidence.

Introduction

Interactive case simulation tool (ICST) is an important vehicle to disseminate clinical evidence. The New York State HIV-HCV-STD Clinical Education Initiative (CEI) has developed various guideline-driven ICSTs and disseminated them to thousands of clinicians. User engagement is critical for effective dissemination of online resources. Here we report a study to assess the engagement of clinicians through analyses of ICST usage in the CEI program. The findings from this study will guide future development of online resources for more effective knowledge dissemination.

Methods

We selected five popular ICSTs from the CEI program for analyses: HIV Pre-Exposure Prophylaxis, Anal Dysplasia and Cancer, Substance Use, HIV Exposed Infants, and HIV Testing. Each ICST has three main components: (1) recommendation, which is presented as hyperlinked text; (2) my case, which is user-defined patient case through an interactive, algorithm-based process to explore different options for patient management; and (3) sample case, which is similar to (2), except that the patient case is pre-defined and the decision algorithm is presented as a whole without step-by-step user interactions. We tracked all user interactions with the ICSTs, such as click of a hyperlink and hit of a button, and recorded the usage data during a study period from October 2013 to November 2015. To assess user engagement, we leveraged a network analysis approach to obtain the number of disengaged users during interactions with ICST and the specific location and context when a user left. For statistical analyses, we compared the proportions of disengaged users among the three components of ICST, i.e., recommendation, sample case, and my case. We used Chi-square test, with Fisher’s exact correction when applicable, to examine the significance of the differences.

Results

<table>
<thead>
<tr>
<th></th>
<th>HIV Exposed Infant</th>
<th>Anal Dysplasia and Cancer</th>
<th>HIV Testing</th>
<th>HIV Pre-Exposure Prophylaxis</th>
<th>Substance Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>disengaged/total – My Case</td>
<td>9/349</td>
<td>5/496</td>
<td>22/8546</td>
<td>16/688</td>
<td>4/259</td>
</tr>
<tr>
<td>disengaged/total – Sample Case</td>
<td>1/81</td>
<td>2/85</td>
<td>11/2240</td>
<td>7/104</td>
<td>1/19</td>
</tr>
</tbody>
</table>

p-value – Recommendation vs. My Case: 0.0013 ↑, 0.2135 ↑, 1.0000, 0.0048 ↑, 0.0002 ↑
p-value – Sample Case vs. My Case: 0.6956, 0.2728 ↑, 0.0747 ↑, 0.0126 ↑, 0.0300 ↑
p-value – Recommendation vs. Sample: 0.0159 ↑, 1.0000, 0.4028 ↑, 0.9173, 0.6843

↑ indicates that the difference is in a direction consistent with the hypothesis. Bold font: p-value < 0.05.

Conclusion

This study provides empirical evidence on impact of usability design to user engagement. Interactive case simulations likely could engage users more effectively when compared to general recommendations delivered in hyperlinked text. User-defined, interactive case simulations could engage user slightly better than pre-defined, static cases. No significant differences found between pre-defined, static cases and general recommendations. Future research is required to examine the generalizability of the findings to other domains and problems.
Automatic Identification of Social Determinants of Health from Clinical Records

Kevin Lybarger1, Mari Ostendorf, PhD1, Ozlem Uzuner, PhD3, Meliha Yetisgen, PhD2
1Department of Electrical Engineering, 2Department of Biomedical and Health Informatics, University of Washington, Seattle, WA
3Department of Information Sciences and Technology, George Mason University, Fairfax, VA

Abstract
Social determinants of health (SDOH) negatively impact morbidity and mortality especially in marginalized and vulnerable populations. For clinical care, it has long been practice to record SDOH as part of social history during clinical care as this history impacts not only diagnosis but also treatment options. SDOH information can improve public health through secondary use applications, like clinical decision-support systems; however, SDOH must first be automatically extracted from unstructured clinical narrative text. In this work, we apply automatic SDOH extraction methods to two publicly available datasets, namely i2b2 Obesity Challenge dataset [1] and MIMIC-3 dataset [2], and present prevalence of SDOH information in these corpora.

Introduction
SDOH focuses on social factors related to health. In our work, these factors include substance abuse (alcohol, drug, tobacco), living situation, and occupation information. Each of these determinants is characterized through status predictions and multi-word spans. Determinant status has the possible values of unknown, none, current, or past. Substance abuse is characterized through the entities: amount (e.g. “minimal” or “three drinks”), frequency (e.g. “daily” or “occasionally”), exposure history (e.g. “since 1990” or “more than 35 years”), quit history (e.g. “four years ago” or “in 1985”), and type (e.g. “wine” or “illicit”). The living situation entity is method (e.g. “lives with her husband” or “lives alone”). The occupation entities include type (e.g. “carpenter” or “running a day care”), method (e.g. “unemployed” or “retired on disability”), and location (e.g. “Florida” or “XYZ Medical Center”).

Methods
We trained a multi-task model SDOH from clinical text. Details of the multi-task model based on bi-LSTM can be found in [3]. To train our model we utilized our publicly available corpus of 364 social history sections (YV-Notes) extracted from 516 history and physical notes downloaded from the MTSamples website [4]. We applied the trained model to 61,472 social history sections from 59,652 discharge summaries and 141,624 physician notes from MIMIC-3 dataset. We also applied the trained model to 472 social history sections from 1237 notes from i2b2 obesity challenge (both training and test sets).

Results
Extraction performance of our multi-task model on YV-Notes is presented in Table 1. We identified SDOH information in both i2b2 and MIMIC-3 datasets. The resulting statistics are provided in Table 2.

Conclusion
The social history section in clinical text indeed contains a wealth of SDOH information, which can be used in both clinical care and in clinical research. The extraction software and automatically generated annotations on MIMIC-3 are released at https://github.com/Lybarger/clinical_extraction. These resources facilitate robust correlation studies between social history factors and disease outcomes. We will manually validate the automatic annotations on the i2b2 obesity challenge to generate a gold standard for a future community challenge.

References

Table 1. Extraction performance on YV-Notes

<table>
<thead>
<tr>
<th>Determinant</th>
<th>Label</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>alcohol</td>
<td>status</td>
<td>0.91</td>
</tr>
<tr>
<td>drug</td>
<td></td>
<td>0.82</td>
</tr>
<tr>
<td>tobacco</td>
<td></td>
<td>0.89</td>
</tr>
<tr>
<td>living situation</td>
<td></td>
<td>0.89</td>
</tr>
<tr>
<td>occupation</td>
<td></td>
<td>0.70</td>
</tr>
<tr>
<td>substance (alcohol, drug, and tobacco)</td>
<td>type</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>amount</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>frequency</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>exposure history</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>quit history</td>
<td>0.86</td>
</tr>
<tr>
<td>living situation</td>
<td>method</td>
<td>0.63</td>
</tr>
<tr>
<td>occupation</td>
<td></td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>location</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 2. Automatic annotation statistics

<table>
<thead>
<tr>
<th>Determinant</th>
<th>I2b2-Obesity Challenge</th>
<th>MIMIC-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>alcohol</td>
<td>341</td>
<td>31,858</td>
</tr>
<tr>
<td>drug</td>
<td>108</td>
<td>16,181</td>
</tr>
<tr>
<td>tobacco</td>
<td>396</td>
<td>35,301</td>
</tr>
<tr>
<td>living situation</td>
<td>216</td>
<td>21,878</td>
</tr>
<tr>
<td>occupation</td>
<td>125</td>
<td>15,486</td>
</tr>
</tbody>
</table>
Value Set Development for Studies Involving Historical Data

Kristin A. Lyman, JD, MHA, Melanie Canterberry, PhD, Elizabeth Crull, MPH, Jason Denton, BS, Fern FitzHenry, PhD, RN, Qiaohong Hu, MS, Sharidan K. Parr, MD, MS, Amy M. Perkins, MS, Dina Zein, MHA, Michael E. Matheny, MD, MS, MPH, Jason N. Doctor, PhD, Daniella Meeker, PhD

1Louisiana Public Health Institute, New Orleans, LA; 2Tennessee Valley Healthcare System, Veteran’s Health Administration, Nashville, TN; 3Vanderbilt University Medical Center, Nashville, TN; 4University of Southern California, Los Angeles, CA

Introduction

Through participation in the National Patient-Centered Clinical Research Network (PCORnet), the REACHnet and pSCANNER clinical research networks were funded to develop the PCORnet Opioid Surveillance Demonstration, a rapid response study that leverages PCORnet Common Data Model data (EHR data) across 5 PCORnet clinical research networks. The overall goal of the study is to create an EHR-based surveillance framework to complement existing CDC and other national and regional surveillance data sources. As an integral part of the study, the study team was required to develop medication, diagnosis, procedure, and lab value sets to be applied when querying historical data from the time period 2010-2017. The goal of this poster is to highlight the challenges and solutions associated with the development of value sets for studies and surveillance models that leverage historical data.

Methods

Value sets were developed for the following: 1) Medications (RxNorm CUI and NDC) - opioids, benzodiazepines, naloxone, buprenorphine, naltrexone, and methadone; 2) Diagnoses (ICD9CM and ICD10CM) - cancer, mental health conditions, substance use disorders, opioid overdose, smoking, HIV, hepatitis B virus, hepatitis C virus, suicide, and neonatal abstinence syndrome; 3) Procedures (CPT, HCPCS, ICD9CM procedure codes, ICD10PCS) - chemotherapy, radiation, and urine drug screening; 4) Labs (LOINC) - urine drug screening.

Several techniques were employed to augment completeness and relevance of value sets for this study. First, multiple published value sets were combined to provide a starting point for further value set augmentation and fine-tuning, including Value Set Authority Center value sets, other published value sets (i.e., those published for endorsed quality measures), Athena OHDSI, and RxNav. Second, current and historical code releases were manually reviewed to ensure that all codes within relevant ranges were captured. Third, in some cases published reports regarding specific historical code changes were reviewed to provide additional validation. Finally, value sets were reviewed manually to ensure that codes align with use of the value set within the larger analytics model that aims to characterize and assess relationships between opioid epidemic risk factors, processes, and outcomes of interest. Any study or surveillance model would benefit from the application of these methods to the development of value sets.

Results

The development of value sets for this study highlighted several challenges that are generalizable to the development of other studies and surveillance models. First, when leveraging value sets for studies involving historical data, value sets must include codes that have been deleted or replaced, including temporary codes. Second, value sets must often be fine-tuned based on a study’s specific aims and population of interest. In developing the value set for mental health diagnoses most often associated with opioid use, misuse, or abuse, it was determined that ICD10CM code R63.2 (Polyphagia) should be deleted. Third, the level of granularity of certain codes may not be appropriate given the level of data present. Naloxone RxNorm CUIs with IN term type were removed due to the potential to include Buprenorphine/Naloxone combinations for sites that had RxNorm CUI coded at the ingredient versus a more granular level. Finally, certain value sets may not include the full scope of codes applicable to a certain concept. A number of published value sets for HIV did not include the ICD9CM code 044.9 (HIV infection, unspecified).

Conclusions

The challenges encountered in the development of value sets for the PCORnet Opioid Surveillance Demonstration are not unique to the opioid use case. Methods can be easily generalized to any study or surveillance model that leverages historical data and requires tailoring of value sets to specific research questions and/or populations of interest.
Scalable Dimensionality Reduction in Patient Medication Domain: A Use Case for ATC Terminology Mapping for Machine Learning Models

Sina Madani, MD, PhD, Colin G. Walsh, MD, MA, Joseph A. Huenecke, PharmD, Jeff Byrd, Asli Weitkamp, PhD
Vanderbilt University Medical Center, Department of HealthIT

Introduction

Machine learning models classify or predict an output likelihood or class from input data. Recent machine learning applications show great potential in patient care where hundreds, if not thousands, of predictor variables co-exist together. However, input data like medications usually contain highly dimensional attributes that may negatively affect algorithm performance or model fit. Patient medication modules within Electronic Medical Record (EMR) systems often identify a drug both at the drug name and dosage/route levels, while, at the latter level, dosage and route of administration information are also included in the name. For example, Clindamycin, Clindamycin 150 mg capsule, Clindamycin 300 mg capsule, and Clindamycin 1% Lotion, are typically represented with four unique identifiers where the last three with dosages/routes information are linked to the parent drug name Clindamycin. The legacy medication authoritative repository at Vanderbilt University Medical Center (VUMC) captured patient “current medications” as part of patient summary services. We are operationalizing multiple machine learning algorithms at VUMC such as suicide risk prediction which rely on legacy data. The Anatomical Therapeutic Chemical (ATC) Medication Terminology has been shown to be effective in dimensionality reduction for suicide attempt risk prediction. ATC translates dosage/route level to higher levels of abstraction. While medication names in the legacy repository show minimum variations, the difference in their dosage and route of administration levels are noticeable among prescribed medications. We hypothesized that by leveraging existing crosswalks and medication classifications in our newly migrated EMR system, we could translate historical information entered in the legacy EMR to the ATC terminology and reduce the dimensionality for clinical prediction in a scalable way while supporting the translation of published suicide attempt risk predictive model to clinical practice.

Methods

We have identified all instances of legacy patient medications (n=15.6 million) between 2008-2017 from 1.8 million patients at VUMC. This dataset corresponds to 61,809 unique medication codes at drug name and dosage/route levels. We leveraged National Drug Data File Plus (NDDF) coding schema to crosswalk from our legacy prescribed medication codes to ATC Terminology (v2018). If multiple mappings existed for a given medication, we aggregated all the ATC codes for that drug. For example, based on the crosswalk, Salicylic Acid would map to three ATC codes representing Salicylic Acid Preparations, Wart and Anti-Corn Preparations, and Other concepts in ATC terminology.

Results

Most medication instances in our legacy dataset were mapped to 2,357 ATC codes (Figure 1). All medications at the dosage/route level were mapped to only 1 ATC code (if there were any), while 3% of the medications at the name level were mapped to more than 1 ATC code (n=1,985). Despite the fact that some of the drugs in the unmapped set had an NDDF Plus code assigned to them at the name (22%) and dosage (9%) levels, no equivalent ATC code was available for such medications. For example, Lutein with a valid NDDF Plus code has no equivalent ATC code in the crosswalk (v2018) that we used for this analysis. We intend to apply pattern matching algorithms in future to the unmapped corpora to find the nearest ATC concept.

Discussions

This study validates the feasibility of using NDDF Plus schema for reliable mapping of legacy patient medications to ATC codes for drug classification in an operational environment for a machine learning algorithm designed for predicting risk of suicide attempts. We will also use this framework in applications such as clinical decision support systems, that rely on higher level of abstractions represented in a standard fashion, to create sharable and reusable knowledge objects within the healthcare enterprise.
Drug-drug Interaction Extraction via Transfer Learning

Yuqing Mao, PhD¹, Kin Wah Fung, MD¹, Dina Demner-Fushman, MD, PhD¹
¹Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Introduction

Drug-drug interaction (DDI) is an unexpected modification in the effect of a drug when taken in combination with another drug. It has the potential to cause significant harm to the patient. The U.S. Food and Drug Administration (FDA) and the National Library of Medicine (NLM) have worked collaboratively on transforming the content of Structured Product Labeling (SPL) documents for prescription drugs into machine-readable data. To inform future FDA efforts at automating important safety processes, the Text Analysis Conference (TAC) 2018 DDI track¹ has provided unique datasets for testing the performance on extracting DDI information from SPL documents of various natural language processing (NLP) approaches.

Recently, deep learning approaches have made significant progress in many NLP tasks. The most important advantage of deep learning approaches is that they do not need manually defined lexical features. More breakthroughs in NLP area were achieved via transfer learning in last year. Transfer learning for NLP tasks includes two steps: the first step is to pre-train a model on large amounts of unlabeled texts; the second step is to transfer the learned general language knowledge to a specific NLP task through fine-tuning. Pre-trained model can improve performances on NLP tasks also because the context can be encoded in its textural representations. In this work, we focus on building a transfer learning framework with BERT², one of the best publicly available, large-scale, pre-trained models, which we used for extracting DDI information from SPL documents.

Methods and Results

The TAC 2018 DDI track includes four shared tasks, and our model is designed to tackle the first two tasks. Task 1 is to extract mentions of interacting drugs/substances, interaction triggers and specific interactions at sentence level. Task 2 is to identify interactions, including the interacting drugs, the specific interaction types: pharmacodynamics (PD) pharmacokinetic (PK), or unspecified, and the outcomes of PD and PK interactions. Since the second task requires the output of the first task, we model them jointly by encoding the associated interaction type for each precipitant drug. Our framework contains three major components: an encoder for contextual embeddings, a BiLSTM-CRF to recognize all entities including precipitants, triggers and effects, and a CNN with two separate dense output layers (one binary classifier for PD interactions and one multiclass classifier for PK interactions) to predict outcome. We utilize the pre-trained BERT model as the context encoder, and compare its performance with one of the state-of-the-art models in TAC 2018 DDI track, which has the same architecture but uses a Bi-LSTM encoder that composes a context representation and character-CNN composed representations. The results of evaluation on TAC 2018 DDI track datasets are shown in Table 1. It is clear that the integration of the pre-trained BERT model can improve the performance of DDI extraction tasks.

Table 1. F1-scores of our BERT-based model and the SOTA model based on Bi-LSTM + CNN.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Set 1, Task 1</th>
<th>Test Set 1, Task 2</th>
<th>Test Set 2, Task 1</th>
<th>Test Set 2, Task 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-LSTM + CNN</td>
<td>33.00</td>
<td>21.59</td>
<td>38.28</td>
<td>23.55</td>
</tr>
<tr>
<td>BERT_BASE</td>
<td>35.58</td>
<td>24.32</td>
<td>39.94</td>
<td>25.76</td>
</tr>
<tr>
<td>BERT_LARGE</td>
<td>41.03</td>
<td>28.10</td>
<td>44.76</td>
<td>30.93</td>
</tr>
</tbody>
</table>

Conclusion

Transfer learning shows improvements in DDI extraction tasks. The performance can potentially be further improved if the pre-trained model could learn biomedical domain specific knowledge through training with a large set of unlabeled texts such as SPL documents and PubMed Central full-text articles.

References

The My Diabetes Dashboard: Preliminary Results of a Prospective, Longitudinal Usability Study
William Martinez, MD, MS; Gerald Hickson, MD; Kenneth Wallston, PhD; S. Trent Rosenbloom, MD, MPH; Tom Elasy, MD, MPH
Vanderbilt University Medical Center - Nashville, TN

Background
We recently applied user-centered design sprint methodology and key strategies for patient engagement to develop a patient web portal intervention for patients with diabetes called the My Diabetes Dashboard (MyDD). It was designed to help patients better understand their diabetes health data as well as support self-management. The dashboard uses graphics to visualize and summarize patients' diabetes health data, incorporates motivational strategies (e.g., social comparisons and gamification), and provides literacy-level sensitive educational resources.

Objective
This study aimed to assess the acceptability and usability of the MyDD over time among patients with type 2 diabetes mellitus (T2DM) as well as users’ attitudes about specific features and potential improvements (e.g., additional functionality) to help patients better understand their diabetes health data and support self-management.

Methods
To accomplish this objective, we conducted a prospective, longitudinal usability study. We recruited 60 adult patients with T2DM who completed the study. Patients were recruited from the Vanderbilt University Medical Center Adult Primary Care (VAPC) clinic and Eskind Diabetes Clinic both located in Nashville, TN. Participants were given one month of access to the My Diabetes Dashboard embedded within Vanderbilt’s patient web portal, My Health at Vanderbilt, run on Epic’s MyChart platform. Data were collected on dashboard usage. At the conclusion of the study, participants completed a questionnaire including a validated measure of usability, System Usability Scale (SUS), and questions about their experience using the MyDD including attitudes about specific features and potential improvements.

Results
Participants’ mean age was 58 (33% were 65 or older), 55% were female, 72% were Caucasian, 18% had limited health literacy, 40% were taking insulin, and the average hemoglobin A1c level was 7.5% (range: 4.9 to 12.5). All participants reported they intended to continue to use the dashboard going forward. The majority of participants (80%) visited the dashboard 3 or more times and 50% spent a total of 15 minutes or more on the dashboard during the one-month study period. Participants’ mean SUS score was 77.9 (SD ±15.2) and significantly greater than the threshold value of 68 indicative of ‘above average’ usability (P<0.001). Participants most commonly identified MyDD’s literacy-level sensitive educational resources as a feature that contributed to improved understanding of their diabetes health data as well as those that would make the dashboard more useful (65% and 73%, respectively).

Discussion
Our study describes the usability of a patient portal intervention designed to help patients understand and manage their diabetes. All participants found the dashboard acceptable (i.e., intended to continue using it). Participants’ SUS scores indicated significantly ‘above average’ usability. Participants identified key features that contributed to improved understanding of their diabetes health data as well as those that would increase MyDD’s usefulness.

Acknowledgements
This work was supported by NIDDK/NIH (K23DK106511 and 2P30DK092986) and NCATS/NIH UL1 TR000445.

References
A Review of ER Case Reports for Potential Anchoring Bias

Roni Matin, MSc1; Amy Franklin, PhD1

1University of Texas Health Science Center, School of Biomedical Informatics, Houston, Texas

Introduction
Research into sociotechnical systems in healthcare indicate the majority of adverse events involve either system errors, cognitive errors or both.1–3 and medical errors due to faulty diagnostic reasoning frequently involve cognitive factors.4 Decision making is widely considered to be a dual process model of system one processes, which are fast, intuitive, often termed heuristics, and system two processes, which are slower, deliberate and effortful. Under time pressure clinicians may defer to heuristics for making diagnoses. While heuristics typically serve the decision maker well, they represent a source of cognitive bias. One such bias identified as impacting clinical reasoning is anchoring bias, which is described by Croskerry as “the tendency to lock onto the salient features of a case too early in the diagnostic process” and failing to adjust from this view.5,6 Emergency departments represent pressured, stressful settings, where clinicians are required to provide prompt diagnosis and treatment to patients, many with life threatening conditions. Additionally, they may not have timely access to all the medical information pertaining to the patient under their care. Unsurprisingly, 65% of ED malpractice claims involve diagnostic error.7,8 Some diagnoses may involve transitions of care and the use of handoff protocols is not consistently universal. Here we reviewed a set of ED case reports involving adverse patient events due to diagnostic error, in order to infer whether anchoring bias may have contributed to the incorrect diagnosis and thereby discover gaps for further study.

Methods
A review of a limited set of case reports was conducted from articles retrieved by ad hoc manual searches in PubMed, and Google Scholar using the search terms “case reports” AND (“missed diagnosis” OR “delayed diagnosis”). Article titles and abstracts were reviewed and ten articles related to cases of missed or delayed diagnosis of patients presenting in ER settings, were selected for full text review and analysis. Case reports where an error had occurred and was caught late in the process or not all resulting in patient harm, were selected and reviewed to note diagnosis and follow on actions reported for each case. Data regarding the patient presentation, the tests, treatment and medications the patient was given following the initial diagnosis, as well as the final diagnosis reported, what prompted the final diagnosis and the case outcome were noted.

Results
The articles (n=10) were from case reports of self-declared diagnostic failures published between 2005-2019 were retrieved and reviewed. Of these, four cases resulted in the death of the patient, four resulted in delayed surgeries and two had outcomes where the patient was left severely disabled. Half the cases involved senior patients and six involved patients with multiple pre-existing chronic conditions such as diabetes, renal failure and HIV. In eight of the ten cases the error in diagnosis was not caught in the initial ED admission and the final diagnosis was reached following tests post transfer to other units or on readmission to the ED. These failures were due to anchoring on patients’ pre-existing conditions or erroneous patient-provided medical history, despite the presence of clinical evidence suggesting otherwise. Also, in eight cases the treatment given was consistent with the initial diagnosis. In two cases, the diagnosis and treatment of the patient occurred over ED shift changes and the authors suggest failures in the handoff practices promoted anchoring bias in the diagnosis.

Conclusion
The articles were heterogeneous in terms type of information and the level of detail reported regarding exact timings, the decision makers involved and the content of communications. Consequently, the findings can only infer the occurrence of anchoring bias involvement in these cases. Of particular interest were the two cases where issues in ED handoffs were identified as contributing to anchoring bias. These two studies contained very detailed reporting of the cases enabling better reflection of the events that occurred. Anchoring effect is recognized to be involved in ED diagnosis errors and these findings suggest may be more prevalent with patients who are elderly or already have complex chronic conditions. The manual selection of the articles does bias the results and limits the generalizability of the findings of this review. However, the findings do suggest further study of the anchoring bias in ED handoffs presents an opportunity of future research to inform the development of handover tools and de-biasing strategies for both medical training and practice.

References
Acceptability of Text Messaging to Help African American Women Manage Anxiety and Depression

Terika McCall, MPH, MBA¹, Todd Schwartz, DrPH²,³, Saif Khairat, PhD, MPH¹,²

¹Carolina Health Informatics Program, University of North Carolina at Chapel Hill, Chapel Hill, NC; ²School of Nursing, University of North Carolina at Chapel Hill, Chapel Hill, NC; ³Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC

Introduction

The rates of mental illness among African American women are comparable to the general population (18.6% vs. 18.9%), however they significantly underutilize mental health services compared to their white counterparts (10.6% vs. 23.4%) (1). Previous studies revealed that telehealth interventions are effective, and can be used to increase access to services (2). The aim of this exploratory study is to gauge the acceptability of utilizing text messaging to deliver mental health services to help African American women manage anxiety and depression.

Methods

A self-administered web-based survey was launched in June and closed August 2018. Eligible participants were African American women (≥ 18 years of age), regardless of mental health history. Participants were recruited through convenience sampling methods (including snowball sampling), via email or social media posts. Respondents were provided an anonymous link to the online questionnaire. The survey consisted of 53 questions, and the following subjects: sociodemographic characteristics, attitudes toward seeking professional psychological help (3), mobile phone use, and acceptability of using a mobile phone to receive mental health care.

Results

The results of this exploratory study (N=101, 90% completion rate) showed low acceptance of the use of text messaging to communicate with a professional to receive help to manage anxiety and depression (< 50% endorsed). Fifty-one percent agreed that having the option to use text messaging to communicate with a professional if they are dealing with anxiety would be helpful. Similarly, 49% agreed that having the option to use text messaging to communicate with a professional if they are dealing with depression would be helpful. Over 50% of respondents indicated having concerns about using text messaging to communicate with a professional. No statistically significant associations were found between age and agreement with the use of text messaging (all p > .05).

Conclusion

The use of text messaging was not highly endorsed by African American women as an acceptable mode of communication with a professional to help them manage anxiety or depression. Concerns around privacy, confidentiality, and the impersonal feel of communicating about sensitive issues via text messages must be addressed in order for this mode of communication to be a viable option. The findings of this study demonstrated the need for additional research into the use of mobile technology to provide African American women with more accessible and convenient options for mental health care.

References


Acknowledgements

The first author is supported by the National Library of Medicine under the Institutional Training Grant for Research Training in Biomedical Informatics and Data Science [T15-LM012500].
Nurses’ SNOMED CT and Physicians’ SNOMED CT have little overlap in terms

Smruti Mehta, BS, Miguel Colon, BSN, Zachary Warren, BS, Karen Dunn Lopez, PhD, Andrew D. Boyd, MD
Department of Biomedical and Health Information Sciences, College of Applied Health Sciences, University of Illinois at Chicago, Chicago IL USA

Introduction:
The objective of this study was to measure the discrepancies within the interdisciplinary documentation of hospital staff within the Electronic Health Record (EHR). The electronic health records allow health information to be electronically shared and viewed by all the members of the healthcare team; however, many times different members will use different terms to describe the same care, which can lead to confusion⁠¹ between the team. This study evaluates the terminology between nurses and physician in patient documentation using SNOMED CT.

Methods:
We analyzed 10 heart failure patient discharge summary notes that were transformed to SNOMED CT using Natural Language Processing program MedLEE¹. Nursing documentation was coded¹ in NANDA-I, NOC, and NIC in HANDS© (a nursing documentation software) and then translated to SNOMED CT². SNOMED CT terms were labeled by provider if the term was used by a nurse or a physician. The relationships between SNOMED CT terms were extracted from SNOMED CT². The relationships between the terms were graphed using Cytoscape 3.6.1 (https://cytoscape.org/) to evaluate similarity and distance apart. In SNOMED CT all terms are seven nodes apart from the root term, so to evaluate similarity in this study if the nurse and physician terms are two nodes or less apart than they are considered to be similar between the professions.

Results:
From the 10 patient discharge summaries, around 15.6% of terms did not map to SNOMED CT. Out of over 300 terms, only two SNOMED CT² terms (pain and cardiology procedures) overlapped within two nodes between the professions. In figure 1, a graph represents interdisciplinary documentation terms of patient care from the EHR. In the graph, the red nodes represent nurse terms, and blue nodes physician terms. All other patients have similar sparse branching patterns.

Discussion:
Based on the results, physicians and nurses used different terms to record the care on the same patient. Although SNOMED CT is a universal health terminology taxonomy, the provider terms were in completely different areas of the SNOMED CT hierarchy². The structural difference represents the fact that physicians and nurses have different scopes of care. The use of different terms even within the same terminology can lead to misunderstanding amongst the caregivers. Within UMLS’s network on average 4 terms per patient were within 2 terms¹, in this study the average was approximately 0.

Conclusion:
From the 10 patients discharge summaries only two SNOMED CT² terms (pain and cardiology procedures) overlapped within two nodes between the professions. The discontinuity between the two professions can be seen through these results. While continuing with this study, we hope to create a better understanding between physician and nurse's documentation even when using the same terminology.

References:

Funding: This project was funded by the National Cancer Institute R01 CA225446-01 SCH: INT: MyPHA: Automatically generating personalized accounts of in-patient hospitalizations.
An Analysis of Advance Directive Documentation at Memorial Sloan Kettering Cancer Center

Chelsea L. Michael\textsuperscript{1,2}, Anjali V. Desai, MD\textsuperscript{1,2}, Gilad J. Kuperman, MD, PhD\textsuperscript{1}
\textsuperscript{1}Memorial Sloan Kettering Cancer Center, New York, NY; \textsuperscript{2}Weill Cornell Graduate School of Medical Sciences, New York, NY

Introduction

Incorporating patient values into treatment planning is critical for the delivery of high-quality care. Important aspects of the patient’s decision-making preferences are highlighted in advance directive documents (ADD). Making ADD easier to access can help ensure that clinical decisions are aligned with the patient’s wishes. As part of an effort to facilitate goal-concordant care by improving usability of ADD and other values-related data in the electronic health record (EHR) [Desai, A V. J Oncol Pract. 2018;14(12):e775-e785], we sought to understand which types of ADD are present in Memorial Sloan Kettering Cancer Center (MSK)’s EHR.

Methods

MSK is a large, single dedicated cancer center. For ADD, a patient may use an MSK health care proxy (HCP) form (with optional living will and organ donation sections) or an ADD from an outside source. The care team forwards completed ADD (and other hard copy documents) to a central scanning group within the medical records department. This scanning group abstracts the document execution date and, for outside documents, identifies the document type.

Having obtained IRB approval, we randomly identified 100 patients who had at least one ADD scanned into our EHR within the last year (2017-2018). We then reviewed all ADD in the EHR for those patients. We classified each ADD by source (our institutional form or outside documents), duplicate scans (multiple scans of the same document) and incomplete or combined documents. We identified elements present in each document (i.e., HCP, living will, organ donation). We compared the abstracted execution date to the execution date documented on the scanned form.

Results

Of the 100 EHR charts reviewed, 49 patients had a single ADD scan (scan), and 51 patients had multiple scans. Of 192 scans in total, 151 were unique documents and 41 were duplicates. Of the 192 scans, 8 were missing pages, and 4 contained more than one document in a single scan.

Characteristics of the 151 unique documents are summarized in Table 1. The majority constituted a HCP form, while living will and organ donation elements were included in a minority of forms. All documents were health-related (i.e., no financial-only power of attorney documents), and 45 documents replaced prior ADD (e.g., a change in health care agent, updated contact information and/or married name), 2 documents re-executed prior ADD with errors, and 1 document voided an older document without specifying updated directives. The 42 outside documents had varied formats and sources, including forms from 2 state health departments and 3 non-profit organizations; lawyer-drafted text documents from 3 states; and 24 forms and text documents from unknown sources. For the 192 scans, the abstracted date did not match the document execution date for 67 scans (35%). Most date offsets were short, but some were much longer (median = 7 days, range = 1-6364 days).

<table>
<thead>
<tr>
<th>Table 1. Document Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
</tr>
<tr>
<td>Total Documents</td>
</tr>
<tr>
<td>Unique Documents</td>
</tr>
<tr>
<td>Institutional Form</td>
</tr>
<tr>
<td>External Form or Document</td>
</tr>
<tr>
<td>Document Elements (non-exclusive)</td>
</tr>
<tr>
<td>Health Care Proxy (HCP)</td>
</tr>
<tr>
<td>Living Will</td>
</tr>
<tr>
<td>Organ Donation</td>
</tr>
</tbody>
</table>

Conclusions

The ADD in our EHR contain multiple elements relevant to the patient’s wishes. Currently, we can only identify these important elements through manual review. Challenges related to paper documentation (e.g., missing pages, multiple copies, incorrect abstraction dates) are frequently present. To improve usability, we first aim to facilitate ADD retrieval from the most used EHR components. Anecdotes suggest that low ADD accessibility in the EHR leads to quality issues (e.g., duplicates). We could also leverage the patient portal for ADD upload and/or review. Analyses of ADD workflows and charting are ongoing. There are many opportunities to improve management of advance directive data as we seek to leverage our electronic systems to facilitate the delivery of high-quality, goal-concordant care.
Machine learning for personalized decision support with patient-generated health data

Elliot G. Mitchell, MA\textsuperscript{1}, Lena Mamykina, PhD\textsuperscript{1}, Matthew E. Levine, BA\textsuperscript{2}, Esteban G. Tabak, PhD\textsuperscript{3}, David J. Albers, PhD\textsuperscript{4}

\textsuperscript{1}Columbia University, Department of Biomedical Informatics, New York, NY; \textsuperscript{2}California Institute of Technology, Pasadena, CA; \textsuperscript{3}Courant Institute of Mathematical Sciences, New York, NY; \textsuperscript{4}University of Colorado, Departments of Pediatrics and Informatics, Aurora, CO

Introduction

We all seek to make informed decisions; in complex domains like health, it can be difficult for individuals to anticipate the consequences of daily choices. This is especially true for diseases with high individual variability. Clinical guidelines only apply to the average individual.\textsuperscript{1} Machine learning and patient generated health data can find patterns and offer insights at the individual level.\textsuperscript{2}

Here, we apply attributable components analysis (ACA)\textsuperscript{3} – a recently developed non-parametric density estimation method based on optimal transport theory – to type 2 diabetes (T2D) self-monitoring data. ACA identifies patterns of association between nutrition and changes to blood glucose (BG) after a meal, and we compare the method to linear regression to explore its potential advantages to support individualized decision support.

Methods

We used data for 15 individuals with T2D collected using a custom smartphone application in prior research. The data included meal records and BG readings recorded before and two hours after each meal. The glycemic impact of a meal was calculated as the difference between post- and pre-meal BG. Each meal was evaluated by a registered dietitian (RD) using a standardized protocol to assess the calories and macronutrients (carbohydrates, protein, and fat) in grams. Each data set included a median of 67 meals (range: 34 to 293), and participants were an average of 54 years old (SD=7.8) with a body mass index of 36.6 (SD=11.9). Analysis with ACA and linear regression were performed on each individual patient’s data set, as well as subsets by meal type (breakfast, lunch, and dinner) for a total of 60 (4 * 15) n-of-1 experiments. To estimate uncertainty and generate confidence intervals (CIs), we used a bootstrap procedure to run each method multiple times on random subsamples of the data.\textsuperscript{4} To compare the performance of the two methods, we produced a series of plots examining the shape of the relationship between each individual macronutrient and BG impact. To measure the quality of the CIs, we calculated the percentage of data points that fell within the CI.

Results and Conclusion

ACA was more robust to outliers, was able to find non-linear relationships, and importantly provided broader, more informative confidence intervals. As shown in Table 1, a much larger percentage of data points fell within the CI for ACA than linear regression. Because uncertainty is intrinsic to the practice of medicine, making ML useful for clinical decision support necessitates accurate and informative uncertainty estimates.\textsuperscript{5} Our results demonstrate the promise of ACA for uncertainty-aware machine learning with patient-generated data.

References

Three Rs of Clinical Data Quality – Relevance, Robustness, Repercussions

Sandra H. Mitchell, RPh, MSIS, FASHP\textsuperscript{1}, Gay E. Stahr, BS\textsuperscript{1}, Todd E. Turner\textsuperscript{2}

\textsuperscript{1}US Department of Veterans Health Information Exchange, Contractor – J P Systems, Inc., Clifton, VA; \textsuperscript{2}US Department of Veterans Health Information Exchange, Office Health Informatics/Clinical Informatics Data Management Office, Washington, DC

Introduction

Healthcare interoperability brings new challenges to your EHR data quality. Exchanged data reflects data quality programs across enterprises. When the data is shared, wide variances in adherence to standards means data content is in flux. Thus, leading to challenges in data ingestion/transformation in the data not being usable in the receiving applications.

Three years ago, the Veterans Health Information Exchange (VHIE) pioneered a clinician viewpoint analysis of HL7 message exchanges with the end goal of increasing native ingestion of external data. From a completeness and content perspective the results of the analysis, where they even exist, can be overwhelming. Clinical content improvement comes from focusing on Relevance, Robustness, and Repercussions.

Relevance

Clinicians expect content appropriate to their needs from external sources, but often experience misplaced, miscoded, and missing content.

“To be useful, EHR data from disparate sources need to be aggregated, mapped and harmonized into a single collaborative environment to ensure full visibility of the patient population.”\textsuperscript{1}

Robustness

Robustness refers to what makes data trustworthy, including accuracy, completeness, timeliness, and accessibility. Loss of the “clinical intention”, due to missing, misplaced or miscoded data can make the clinician spend more time with the technology than the patient or can lead to the clinician missing key patient clinical issues.

Repercussions

Data is a business resource and needs to be valued as such. Poor data can result in duplicate testing, incomplete research projects, and impact bringing new drugs to market. If the underlying data is poor, then all dependent applications are impacted.

Patient safety risk increases if poor data is added to the local Electronic Health Record (EHR) or if data requires transformation into local value sets. Without quality data, initiatives such as Artificial Intelligence (AI) or Clinical Decision Support (CDS) can’t perform to expectations. Complete and accurate data is not only a financial imperative, it is also an ethical imperative (e.g. patient safety, access to protocols based on relevant and robust data).

Conclusion

Opportunities exist to rectify the situation. The enthusiasm of the vendors to understand the challenge and adopt the VHIE Interoperability Data Quality process is leading to improved data exchange today. Without complete and high-quality clinical content exchange, the potential of Clinical Decision Support and Artificial Intelligence cannot be realized, research projects will take longer to complete, clinicians will continue to suffer burnout. Improved clinical content within the healthcare interoperability space is the true key to success.

References

Synthetic Data for Teaching Data Integration in Informatics Graduate Program

Hedyeh Mobahi, MS, Hua Min, PhD, Janusz Wojtusiak, PhD
George Mason University, Fairfax, VA

Introduction
Data integration is an important part of the health data analysis process, and a prerequisite to application of any analytic tools. It is critical for students in health informatics and health data analysis programs to become familiar with the required skills to perform such tasks. In Mason’s health informatics curriculum (MS and PhD), students learn these skills with SQL and Python as part of Health Data Integration course. Students learn increasingly complex approaches to data integration: first simple process in which a master patient index (MPI) exists is covered. Then, students need to search and match patient IDs that use different coding systems and conversions are needed. Finally, students learn and apply statistical matching methods. In all of the exercises, students also learn and apply evaluation methods used to check/confirm successful integration.

One challenge for instructors teaching data integration is to provide the right datasets which are realistic but at the same time do not contain any sensitive identifiable patient information. The approach that guarantees this requirement is use of realistic synthetic data. In the presented work, we algorithmically generated synthetic datasets that resemble real ones, yet do not contain any real patient information.

The presented study had two main goals: 1. Generate a usable realistic synthetic dataset for the education purposes; and 2. Ensure that the data shows the challenges needed to study complexity of data integration.

Methods
Synthetic datasets were created to resemble medical claims and cancer registry. The starting point of the process of generating synthetic data was extraction of real deidentified datasets from two independent sources. The extracted subsets from claims data included demographics, diagnosis (ICD9), procedures (HCPCS) and billing information. Cancer registry data included demographics and cancer diagnosis information only.

Records in the synthetic data were generated by randomly selecting age within ±3 years for each patient. Payments were randomly generated within ±10% of the real patient values. HCPCS and ICD9 codes were generated consistently with respective hierarchies for these codes. HCPCS categorization is broad, hence the following strategy was applied: if the code consisted only of digits, a random code was picked from range [10 codes before, 10 codes after] of the real value, if it starts with a character, a random code was picked from codes starting with the same character, and if a code ends with a character a random code from same category was chosen. To generate ICD9 codes, we have used the standard groupings suggested by CDC. For each code, two different thresholds were used. If the code reached the lower threshold the digits before dot were replaced by using a random code from the same category, and if the higher threshold was reached the digits after dot were replaced by picking one random code from the subset starting with the same digits. Race and gender were selected from original data. Finally, the two generated synthetic datasets were matched using statistical matching based on age, gender, race, and cancer diagnosis. Synthetic patient identifiers were generated that link across both datasets.

Generation of the synthetic data was done using Python 3.6 and PostgreSQL.

Results
Using the described process 1,655,574 synthetic claims were generated for 14,977 patients. The total number of 10,392 distinct ICD9 codes, 8,452 distinct HCPCS codes were used and 1,655,574 payments were generated. In the synthetic cancer registry data 37,198 cases of cancer were generated for 35,191 patients. After statistical matching 5,110 patient records remained. The additional 39,948 unmatched records were left in the data to provide students with realistic scenario in which not all records can be matched.

The final dataset is provided to students in five comma separated files: 1. Claims demographics; 2. Claims diagnoses; 3. Claims procedures and payments; 4. Cancer registry demographics; and 5. Cancer registry diagnoses.

Conclusion
The described process allowed us to create synthetic data that is realistic, yet do not contain any real patient information. Such data can be used in teaching data integration and databases without need of providing special security and data access control. It is important to note that such synthetic data are not intended for more advanced analytics courses including health data mining or machine learning as the data are random beyond the possibility of finding useful patterns. We specifically designed it only for the purpose of patient matching.

References
**HemOnc.org: Evaluation of Information Models for Cancer Therapy Representation**

Zachary H. Moldwin, BA\(^1\), Harry Hochheiser, PhD\(^2\), Jeremy L. Warner, MD MS\(^3\)

\(1\)University of Illinois at Chicago College of Pharmacy, Chicago, IL; \(2\)University of Pittsburgh, Pittsburgh, PA; \(3\)Vanderbilt University, Nashville, TN

**Problem Description**

HemOnc.org\(^1\) is a cancer therapy wiki and knowledge base with many possible applications for clinicians, researchers, and clinical decision support systems. Expansion of its functionality requires that it be readily accessible and maintainable, as well as semantically sound and interoperable with health information systems. Tools must be developed to simply, safely, securely, and consistently edit concepts and relationships. Easily navigable user-facing EHR and CDS applications should have access to well-built and defined information representation models. We compare four information models for maintainability, usability in health and research systems, and compatibility with related ontologies, standards, and terminologies.

**Methodology**

With permission of HemOnc.org LLC, the HemOnc.org content was obtained in HTML format, and portions were transformed into an OWL ontology in JSON format. This was converted to XML and parsed into a SQL Server relational database (HemOncDB). The HemOncDB data model uses a standard normalized multi-table domain architecture, derived by analyzing the knowledge representations in the OWL ontology. Forms were created in C# to allow editing of HemOncDB and to display the relationships between concepts. HemOncDB was then transformed into a Neo4j graph database using a Python script and the py2neo library. Portions of the HemOncDB content were also transformed into several denormalized tables using the OHDSI OMOP concept-relationship data model.\(^2\)

**Evaluation Results**

For semantic modelling in the ontology, excessive pre-coordination has been generally avoided because it could limit the flexibility of the derived products/tools and also overly complicate the model. Importation of ontology identifiers into HemOncDB generated data integrity issues because the ontology relied upon meaningful string identifiers rather than Concept Unique Identifiers (CUIs). For example, in the OWL ontology format, ‘Cyclophosphamide’ appears both as a parent element (containing brand names, etc.) and as a child element of many different regimens, and no CUIs were present to ensure that ‘Cyclophosphamide’ the parent and ‘Cyclophosphamide’ the child were semantically the same. Additionally, duplicate instances of many parent concepts, often with differing case or spelling and with differing child elements, were found.

Population of the Neo4j and the OMOP OHDSI databases were relatively straightforward because they are close in structure to the original OWL export format; design of a custom domain model was not necessary. In contrast, HemOncDB has a domain-specific design with unique identifiers and promotes data integrity, concept uniqueness and referential integrity. Therefore, HemOncDB was useful for data cleanup from the original ontology, such as detection of concept duplication. For the non-domain model architectures (OWL ontology, Neo4j and OHDSI OMOP), n-ary relationships were difficult to represent and/or enforce. As an example of a relationship that is difficult to represent/enforce, a treatment regimen used in one disease subtype may be supported by evidence from one publication, whereas the same treatment regimen used in a different disease subtype may be supported by another publication.

**Conclusion**

The HemOncDB data integrity features readily detected OWL dataset irregularities that were not detected through the non-relational models. The Neo4j model provided useful graph traversal and visualization tools that were not readily available in HemOncDB. Representation of regimen data in HemOncDB, OWL, Neo4j and OHDSI formats will increase their availability for oncology research and may allow easier integration with EHR and clinical decision support systems.

**References**

2. Observational Health Data Sciences and Informatics. [Internet]. c2019 [cited 2019 Mar 13]. Welcome to OHDSI. Available from: [https://ohdsi.org/](https://ohdsi.org/)
A Comparison of Online Consumer Medication Information from Canada, Denmark, and the United States of America

Helen Monkman, PhD¹, Thomas Schmidt, PhD², Christian Nohr, PhD²
¹School of Health Information Science, University of Victoria, BC, Canada; ²Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Denmark

Introduction

Many people around the world take prescription medications: some for short durations (e.g., antibiotics) others continuously on a daily or multiple times daily basis. Taking any prescription medication inherently has both benefits (i.e., alleviating symptoms) and risks (e.g., side effects, contraindications, interactions, overdoses) for consumers. This study compared the content and design of online Consumer Medication Information (CMI) from three sources: one Canadian and one American pharmacy, as well as the Danish national resource minmedicin.dk. The investigators identified strengths and weaknesses of each of the three online CMI sources revealing opportunities for improvement.

Methods

We independently analyzed online CMI from three countries (Denmark, Canada, United States) for three medications (an inhaler, a tablet, and an intravenous medication), for a total of nine CMI. We assessed the content of each CMI against Monkman and Kushniruk’s¹¹ content utility criteria to assess whether important topics were present. We also assessed the design of the three online CMI resources by critically examining how each one presented information and how users interacted with the site.

Results

Content: Generally, all nine online CMI addressed the majority of the content criteria. The American CMI was the most comprehensive and the least comprehensive was the Danish CMI. However, each country’s CMI had its shortcomings. We found no drug interaction information in the American CMI, no information concerning missed doses in Danish CMI, and no overdose information in the Canadian CMI. Directions for use varied considerably from very brief to elaborate descriptions. Only the American CMI provided patient to patient medication information using summary statistics drawn from PatientsLikeMe. Finally, only Denmark provided specific information about the frequency of experiencing side effects (e.g., <10 in 100 people). Design: Each online CMI source relied on content structuring using templates. Navigation varied between the three CMI sources. Canada used in-page anchor links, with all content visible on page load. In contrast, both the USA and Denmark leveraged expandable text boxes. When expanded, the American text boxes were lengthy blocks of text. Both the American and Danish sites displayed pictures of the medications. However, only Denmark offered used slideshows and videos to convey directions for use.

Discussion and Conclusion

This analysis of online CMI from three countries, highlighted interesting differences in both the content and design of online CMI. Research dictates that online text should be concise and easily scannable² and pages should not be too long.² However, American and Canadian online CMI violated these guidelines which impedes their usability and likely use as well. Both Canada and Denmark equipped consumers with actionable information to empower consumers, whereas the USA prompted them to consult their healthcare providers. Yet, only the USA provided peer experiences. Denmark’s online CMI demonstrated unique benefits, which may be associated with its streamlined single national CMI approach, rather than relying on pharmacies to provide their own. In summary, each country’s online CMI demonstrated unique strengths and weaknesses and we found opportunities for improvement.

References

Adaptation of a Natural Language Processing Algorithm Following Implementation of a New Electronic Health Record

Sungrim Moon PhD, Vinod Kaggal, Sungwan Sohn PhD, Hongfang Liu PhD, Rajeev Chaudhry MD, Adelaide M. Arruda-Olson MD, PhD
Mayo Clinic, Rochester, MN, USA

Introduction: A rule-based natural language processing (NLP) algorithm has been developed to identify patients with peripheral artery disease (PAD). Subsequently, the NLP algorithm was delivered via the institutional near real-time NLP infrastructure for extraction of relevant patient-specific knowledge from clinical narratives. However, the institution migrated to a new Electronic health record (EHR) system and the PAD-NLP algorithm lost functionality. This study describes the process of adaptation of the PAD-NLP algorithm to restore performance following deployment of the new EHR.

Methods: Two separate versions of the PAD-NLP algorithm using MedTagger software were deployed to process clinical narratives of the new EHR system at Mayo Clinic. The first algorithm was the original PAD-NLP version and the second was a revised version which was modified for use with the new EHR (EPIC). The original PAD-NLP algorithm identified a cohort of 2,574 patients using near real-time NLP infrastructure between October 8th and December 6th, 2018. The algorithm performance was evaluated by manual review of the NLP output for patient classification. The NLP algorithm was revised with updated rules and keywords. "Medical Specialties" and “Note Types” in clinical narrative of the new EHR were mapped to equivalent information in the historical EHR (see Table 1). This revised algorithm was then applied to the same set of clinical narratives. Clinical narratives of 50 randomly selected patients (25 PAD cases and 25 non-cases) identified by the revised algorithm were used for evaluation. These narratives were manual reviewed by two trained annotators following written guidelines for standardization.

Table 1. Mapping of specialties and note types of the 2 EHR systems.

<table>
<thead>
<tr>
<th>EHR system</th>
<th>Medical specialties</th>
<th>Note types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historical EHR</td>
<td>Critical Care; Urgent Care; Primary Care; Hospital Internal Medicine; General Medicine; Family Medicine; Cardiology; Vascular; Vascular Wound Care; Vascular Surgery</td>
<td>Consult; Subsequent Visit; Patient; Progress; Supervisory; Limited Exam; Specialty Evaluation; Multisystem Evaluation</td>
</tr>
<tr>
<td>New EHR</td>
<td>Critical Care Medicine; Community Internal Medicine; General Internal Medicine; Hospital Internal Medicine; Family Medicine; Cardiovascular Diseases; Vascular Surgery; Vascular Medicine</td>
<td>Consults; Consults – Outpatient; Progress Notes; Ambulatory Progress Notes; Ambulatory History and Physical Notes; Interval History and Physical Note; History and Physical</td>
</tr>
</tbody>
</table>

* Underlines represent the mapping between historical EHR and new EHR

Results: In the new EHR lexical variations in clinical narratives were different for representation of the same medical concepts compared to the historical EHR system. After implementation of the new EHR system, there were changes in clinician workflow and clinicians also changed terms used for documentation in the new EHR. For example, the original PAD-NLP algorithm contained the exclusion keywords “deep vein thrombosis” (DVT) or “deep venous thrombosis”. However, in clinical narratives of the new EHR system the concept of DVT was represented as “thrombosis deep vein”, which lead to erroneous exclusions with increase of false-positive results. The revised PAD-NLP algorithm identified only 715 PAD cases (27.78% of 2,574 patients of the cohort identified by the original version of the algorithm. Most of these 2,574 patients were false-positive cases.) In a randomly selected sample (n=50) for evaluation the refined NLP algorithm had accuracy of 98% (95% CI, 89.35%-99.95%), sensitivity of 96% (95% CI, 79.65%-99.90%), specificity of 100% (95% CI, 86.28%-100.00%), positive predictive value of 100% (95% CI, not applicable) and negative predictive value of 96.15% (95% CI, 78.56% to 99.42%).

Conclusions: Migration of EHR systems significantly impacts operation and performance of analytic tools installed in the near real-time NLP infrastructure. This study describes the process to restore functionality and for further refinement of the NLP algorithm mitigating the impact of the EHR migration. Updated rules and keywords as well as mapping of specialties and note types were essential to adapt a rule-based NLP algorithm to the new EHR.

References
Residential neighborhoods and walking: An Informatics Approach to Walkability
Stephen J. Mooney, PhD, Department of Epidemiology, University of Washington, Seattle, WA
Philip Hurvitz, PhD, Urban Form Lab, University of Washington, Seattle, WA
Anne Vernez Moudon, Dr es Sc., Urban Form Lab, University of Washington, Seattle, WA
Brian E Saelens, PhD, Seattle Children’s Research Institute, Seattle, WA

Background and Methods
Walking is the most common form of physical activity among adults. Numerous studies have assessed the potential impacts of assorted neighborhood characteristics on walking. In practice, measureable neighborhood characteristics are often highly inter-correlated. We took an informatics approach to identifying neighborhood environment predictors of objectively measured walking.

We used data from the Travel Assessment and Community (TRAC) study, a 3-wave cohort study of adults living in King County, Washington. In 2007-2009, each adult carried a Global Positioning System (GPS) device and Accelerometer device for 7 days. Using a previously validated algorithm, we identified periods of time the adults spent walking (‘walk bouts’) from these data. We developed the Automatic Context Measurement Tool (ACMT), a piece of R code to compile 146 measures of neighborhood characteristics including population demographics, commute mode, educational attainment, and land cover from publically available data sources with 833 meters (∙½ mile) of each participant’s home address. Following a Neighborhood Environment-Wide Association Study (NE-WAS) design, we used negative binomial regression to model the number of walk bouts each participant completed in 7 days as a function of each neighborhood measure individually, adjusting for selected participant characteristics. Next, we compared predictive accuracy of a model predicting walk bouts using conventional walkability measures to a model predicting walk bouts from the first principal component of the 146 measures.

Results
After Bonferroni correction, many neighborhood characteristics were associated with number of walk bouts (Figure 1). Proportion measures (e.g. proportion of residents commuting by car) were more strongly associated with walking than count measures (e.g. number of residents commuting by car). The model using principal components explained more of the variance in the total number of walk bouts than the models using conventional walkability measures did (Figure 2).

Conclusions
Informatics approaches to identifying residential neighborhood determinants of walking hold promise. Future work will extend the number of measures included, explore other modeling techniques, attempt to replicate findings in other geographic contexts, and explore how results may be of use for public health intervention.

References
Enhancing Accessibility of a Molecular Oncology Almanac for Precision Cancer Medicine

Nicholas S. Moore, BA¹,²,³, Brendan Reardon, BS²,³, Eric Kofman, BS²,³, Nathanael Moore, BS⁴, Eliezer Van Allen, MD²,³
¹Harvard Medical School, Boston, MA; ²Dana-Farber Cancer Institute, Boston, MA; ³Broad Institute of MIT and Harvard, Cambridge, MA; ⁴Indiana University School of Medicine, Indianapolis, IN

Problem
Cancer care is increasingly informed by molecular alterations that predict therapeutic sensitivity, therapeutic resistance, and prognosis. Clinical and preclinical studies continue to identify somatic and germline alterations that impact clinical decision-making, increasing the complexity of interpreting tumor molecular profiles for patient care and research. Increasing access to point-of-care molecular profiling holds great potential for patients but will only translate to improved outcomes if clinicians and researchers are able to interpret patient data in the context of a rapidly changing and broadening research landscape.

Towards solving this problem, we have built and continue to extend the Molecular Oncology Almanac, an open-access combined genomic interpretation algorithm, knowledge system, and web portal for informing treatment decisions through rapid assessment of tumor actionability. Analysis encompasses not only single nucleotide variants, insertions/deletions, and copy number alterations, which are commonly used at the point of care, but also anticipates growing application of fusions, mutational burden, mutational signatures, microsatellite instability, and aneuploidy. A major obstacle to maintaining an up-to-date knowledgebase of actionable alterations lies in the difficulty of incorporating relevant assertions from the literature as soon as they are available. To improve the accessibility of these assertions for the wider oncology research community, we adapted the Molecular Oncology Almanac by implementing open API endpoints and a browser extension for the purpose of rapid assertion acquisition.

Aims
1. Increase accessibility of the Molecular Oncology Almanac precision medicine knowledgebase through open-access API endpoints.
2. Implement a user-friendly interface for submitting actionable alterations to the knowledgebase directly from journal websites.

Methods
RESTful API endpoints for getting and posting alteration-action relationships to the Molecular Oncology Almanac were developed within the existing Flask web framework. Endpoints were designed within the OpenAPI 3.0 standard and documented in Swagger. Using these endpoints, a Google Chrome browser extension for posting alteration-action relationships was developed. Every submitted assertion includes a citation, class of alteration, cancer type, alteration effect, and level of evidence; gene, alteration, and therapy may also be included where relevant. Testing of endpoints and the extension was performed by multiple users in the Van Allen Lab. Submitted assertions are reviewed by site administrators prior to final incorporation into the database.

Results
The Connector browser extension was submitted to the Google Chrome Web Store, with final release pending approval. Multiple test users have submitted assertions through the pre-release extension. The Connector streamlines the process of submitting actionable alterations while reviewing the literature, reducing the effort required for knowledge base maintenance.

Conclusion
Open access tools for posting and retrieving assertions, sources, genes, and specific alterations expand the scope of the Molecular Oncology Almanac for integration into external interpretation resources. These developments represent a step forward in providing accessible clinical interpretation of tumor molecular profiling to clinicians and researchers. The Connector will be presented to cancer research audiences aiming for wide adoption in the community.
Measurement of physician burden in EHRs: a systematic literature review

Amanda J. Moy, MPH1, Sarah Collins Rossetti, RN, PhD1
1Columbia University Department of Biomedical Informatics, New York, NY, USA

Introduction
Public policy discussions associate EHRs with increased documentation burden resulting in burnout. While validated measures of burnout exist (e.g., Maslach Burnout Inventory), burden has largely been ill-defined.

Methods
We used PRISMA guidelines to conduct and report our systematic review. We searched PUBMED from 1/1/1995 to 2/9/2019 for all English-language studies investigating EHR documentation burden among physicians in mixed practice settings. Search terms were: (burden) AND (documentation OR note) OR ehr OR emr) AND (physician OR clinician OR doctor). After removing duplicates, titles/abstracts were independently assessed by two reviewers. Studies were restricted to the U.S. due to its unique socio-political context and healthcare ecosystem. A priori inclusion criteria were: U.S. studies utilizing any documentation burden metric (e.g., perception of time, event logs). A priori exclusion criteria were studies: outside the U.S., not involving physicians, without mention of burden, neither descriptive, observational nor interventional.

Results
Of the 732 articles retrieved, ten studies met criteria. After full-text review, two additional studies were incorporated based on in-text references for a total of 12 studies. Several specialties were examined; half of the studies involved internists and family/primary care. Study designs varied (Table 1), and distinct time constructs were measured: observed time elapsed, “time block[s]” constructed via EHR login data, EHR login/logout time, EHR event logs, self-reported time as a relative metric (i.e., more or less time), self-reported proportion of burden or time, and observed documentation time for notes. We identified six main proxies of burden (Table 1). Physician burnout was mentioned in 41.7% (5/12) studies, but only one used a validated burnout assessment. Only one study distinguished burden from burnout; independently, only one explicitly mentioned generating a quantitative measure of burden.

Table 1.

<table>
<thead>
<tr>
<th>Study design</th>
<th>Study characteristics</th>
<th>Proxies of burden</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Time-and-motion (n=2)</td>
<td>Study Type</td>
<td>1. Number of incomplete charts after shift</td>
</tr>
<tr>
<td>• Retrospective observational (n=1)</td>
<td>Quantitative (n=5); mixed-methods (n=4); qualitative (n=3)</td>
<td>2. Time after shift-hours on computer or clerical tasks</td>
</tr>
<tr>
<td>• Observational field research (n=2)</td>
<td>Site</td>
<td>3. Total EHR login time</td>
</tr>
<tr>
<td>• Dual-balanced crossover design (n=1)</td>
<td>Single (n=8); multi-site (n=4)</td>
<td>4. Proportion of time spent in direct patient care compared to documenting/interacting with the EHR</td>
</tr>
<tr>
<td>• Simulation (n=2), Prospective observational (n=1)</td>
<td>Setting</td>
<td>5. Time spent viewing the EHR</td>
</tr>
<tr>
<td>• Cross-sectional (n=1)</td>
<td>Clinical (n=10); laboratory (n=2)</td>
<td>6. Number of repeated views of EHR data elements within one session</td>
</tr>
<tr>
<td>• Controlled experiment (n=2)</td>
<td>Environment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teaching hospital (n=7); other (n=5)</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion
EHR documentation burden among physicians of diverse specialties remains a challenge in both inpatient and ambulatory settings. Targeting burden as the subject for quality improvement strategies and interventions requires quantifiable measures that are comparable and consistent across time, settings, and contexts—a metric that isolates the problem area, as well as its magnitude and directionality. However, research utilizing concrete, validated measures of burden are lacking. Given the multitudinous definitions for both burden and time, this review suggests evidence is fragmentary. Further research is needed to reliably operationalize the concept of burden, explore how it is best measured, and standardize its utilization.

Acknowledgements: Study supported by the U.S. National Library of Medicine of the NIH under the training fellowship award T15LM00707.

References
Designing a biomedical information system for clinical assessments and subsequent interventions for mental health and substance abuse YMSM patients of color with HIV utilizing Multidimensional Scaling Analysis and Paired Comparisons Techniques

Sukrit Mukherjee, PhD, MSE1, 2, Eric Houston, PhD2

1Center for Biomedical Informatics, Charles R. Drew University of Medicine and Science, Los Angeles, CA, USA; 2Department of Psychiatry and Human Behavior, Charles R. Drew University of Medicine and Science, Los Angeles, CA, USA

Abstract

In mental health and substance abuse treatment, individualized assessments provide information on the specific thoughts and cognitive processes influencing a person’s behavior, emotional responses, and psychological functioning. Given the lack of automated assessment procedures or individualized clinical interventions in the growing health disparities in the South Los Angeles of USA, we developed a novel system using idiographic techniques to automatically and quickly generate individualized patient assessment data for use in clinical interventions.

Introduction

Depression is widespread and disproportionately affects people living with HIV and other chronic health conditions (Ali et al, 2006). A study based on a nationally representative sample of people living with human immunodeficiency virus (HIV) in the United States found that 36% of participants had a probable diagnosis of major depressive disorder, a rate nearly five times greater than that of the general population (Bing, 2001). Receiving treatment for depression becomes critical for those HIV patients but for many people living with HIV (PLH), depression remains largely untreated (Asch, 2003; Shippy, 2004). This low percentage of treatment for depression in the HIV patient population have led to several factors, including the stigma including tendency of individuals to avoid or not to engage in treatment for depression as well as non-participation in depression intervention research (Shippy, 2004; Earnshaw et al, 2013; Nichols, 2002; Reece, 2003). Little research has focused on how to reach and engage patients with depression and HIV who may be at the intersection of multiple social identities faced with stigma. Paired comparisons techniques and MDS analyses are two data analytic techniques that have traditionally been used in statistics and research (Zhu, 2009). But to date, paired-comparisons techniques and MDS have not been combined for use in either automated assessment procedures or individualized clinical interventions. We intended to fill this gap by utilizing novel use of such idiographic approaches that may be particularly valuable in clinical settings that lack staff trained in clinical interviewing or individualized assessment techniques.

Methods & Results

We followed a mixed method research design in this study. The study called, Project STEP (“Steps Towards Embodying Positivity”) was targeted to conceive a new cognitive training intervention system to address depressive symptoms and suboptimal HIV treatment adherence among African American and Latino YMSM (ages 18-29) in the Los Angeles metropolitan area. The system consists of 2 stages. The first stage was designed using Java Server Pages (JSP) as frontend, Java as middle tier and MySQL as backend on a Linux platform with statement session of the input data (paired-comparisons ratings) from a participant and it processes the data using MDS analysis and transforms the data into display of a n-dimensional spatial configuration and provides stress values, a least distance measure and R values of the individual user. The 2nd stage that is designed with ASP.Net as frontend, C# as middle tier and MS SQL Server as backend on a Windows platform receives statements, ratings, dot location and emotion values from first stage and generates stimuli for use in ABM training interventions, particularly individualized ABM interventions. The automated nature of our assessment approach and the relative speed with which the information can be applied is novel and has much potential in addressing gaps in the delivery of care.

Conclusion

We have attempted to solve the problem of clinical assessment for mental health and substance abuse patients belonging to YMSM of color with HIV across the continuum of care by creating a novel system that automatically and quickly generates individualized patient assessment data for use in clinical interventions like attention bias modification training with strong market potential.
Variability in User Response to Custom Alerts in the Electronic Health Record: An Observational Study

Naveen Muthu, MD1, Eric D. Shelov, MD, MBI1, Marc C. Tobias, MD2, Dean J. Karavite, MS1, Evan W. Orenstein, MD3, Robert W. Grundmeier, MD1
1Children’s Hospital of Philadelphia, Philadelphia, PA; 2Phrase Health, Philadelphia, PA; 3Children’s Healthcare of Atlanta, Atlanta, GA

Introduction

Computerized alerts in electronic health record (EHR) systems are intended to bring evidence to clinical practice and improve the safety of patient care. The adoption of EHRs that meet the requirements for the Meaningful Use incentive program has allowed organizations to build custom alerts in the EHR. These custom alerts, as opposed to medication alerts that are often purchased from third-party vendors, remain poorly characterized. Therefore, in this study, we aimed to describe the incidence of and user response to these custom EHR alerts at Children’s Hospital of Philadelphia.

Methods

We examined data for all interruptive custom alerts from October 1, 2017 to December 31, 2017. Non-interruptive alerts were not included. Data were obtained from the institutional clinical data warehouse. One author (N.M.), a pediatrician and board-certified clinical informatician, reviewed the display text of each custom alert and determined if 1) an actionable intervention was available for the user to choose in the alert itself and 2) if an actionable intervention was available, categorized which of the potential alert responses constituted alert acceptance and alert overrides.

Results

137 unique interruptive alerts were displayed on 395,161 occasions to users of the EHR in the study period. Of all the alert instances in the study period, 51.6% were seen by nurses. Prescribers (residents, fellows, attending physicians, nurse practitioners, physician assistants) saw 28.6% of alerts. Pharmacists, who are often targeted by medication prescribing alerts, receive only 7.1% of custom alerts. Of the 137 unique alerts, 70% included an actionable response within the alert itself, and these alerts accounted for 69% of all instances when a user saw an alert in the study period. When an actionable option was offered to the user, users accepted the alert recommendation 11% of the time and overrode the alert 89% of the time. Although most alerts have an acceptance rate of 10% or less, a subset of alerts are accepted more than 30% of the time (Figure 1).

Figure 1. Number of alerts by alert acceptance rate. The majority of alerts have an acceptance rate less than 10%. However, there is a subset of alerts that have an acceptance rate over 30%.

Conclusion

There appears to be a substantial range of alert acceptance for custom alerts, suggesting that despite the potential for alert fatigue, providers consider alert recommendations. Further study of the contextual and display characteristics that explain the variability in provider response to alerts is needed in order to effectively design and use alerts.

References

Improving Automation and Access Using the Vaccine Code Set Management Service

Stuart Myerburg, JD, Centers for Disease Control and Prevention, Atlanta GA
Millie Malai, MPH, Deloitte Consulting, LLP, Atlanta, GA

Problem Addressed:
The use of data standards is not uniform across healthcare providers affecting the quality of information exchanged with public health entities. Specifically, for immunization registries, outdated or incorrect information can affect the accuracy of public health coverage assessments used to make immunization program decisions. This service aims to provide a centralized source for vaccine-related codes to ensure that the mappings between codes, attributes, and metadata are sound, and then offer a web service to facilitate more automated updates.

Description of our service:
The Centers for Disease Control and Prevention (CDC) deployed a centralized service to provide immunization information systems (IIS) and provider EHR systems with comprehensive and consolidated mapping services for vaccine and vaccine-related codes. The codes represent components of vaccine ordering, inventory management, and the documentation of vaccines administered to patients, and are managed by diverse governmental and non-governmental agencies such as CDC (CVX and MVX codes), Food and Drug Administration (NDC codes), and the American Medical Association (CPT codes), amongst others. The centralized service maps CVX, MVX, CPT, NDC (unit of use and unit of sale), Global Trade Identification Numbers (GTIN), vaccine lot numbers (unit of use and unit of sale), and Vaccine Information Statements (VIS) codes, along with metadata associated with the codes. The leading-edge application allows for a more streamlined management of the codes and more automated access to the codes via web service.

The Vaccine Code Set Management Service (VCSMS) was established to explore and implement new tools for management and delivery of these code sets. The tools and data foundations have been implemented for managing and curating catalogs and maps. In April 2019, the service deployed an end-user web service to facilitate more automated use of the information published by CDC.

The main objectives of the service are:
Improving data management and quality – Increase automation to improve process efficiency and quality, establish consistently-applied data editorial policies, and improve sourcing of external reference data used for code set maintenance and mapping.

Expand content structure – Deliver new content data files specifically designed to aid integration with electronic systems.

Provide automated data access – Provide new options for direct delivery of content to end user organizations via web-service and desktop client.

Poster will describe:
The poster will describe the centralized service and how terminologies are managed through the system interface, the changes to the workflow to manage the content, the public health organization’s workflow and how it has changed to improve automation using the service, and lessons learned that other organizations can use when deploying a web service or similar functionality.
Systematic Review of Patient-Facing Visualizations of their Personal Health Data

Annie Myers, MA¹, Samuel Isaac, BS¹, Ruth Masterson Creber, PhD, MSc, RN¹, Meghan Reading Turchioe, PhD, MPH, RN¹

¹Div. of Health Informatics, Dept. of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY

Introduction and Purpose

Tools that allow patients to generate and view health-related data have grown in recent years. In parallel, health-related data is being returned to patients and consumers with increasing frequency, towards the goal of increased engagement, education, and shared decision-making. In the setting of varied literacy, health literacy, and numeracy levels, visualizations may aid patients and consumers to better interpret and act upon health-related data. However, there remains a need to understand best practices for visualizing health data to patients. The purpose of this systematic review was to review and evaluate the state of the science of patient-facing visualizations of their personal health data.

Methods

In December 2018, we searched five scholarly databases (Pubmed, Embase, Scopus, ACM Digital Library, and IEEE computation index) for articles reporting patient-facing visualizations for personal health data, including data originating from both patients (i.e., patient-reported outcomes) and clinical settings (i.e., lab values) published between 2013 and 2018. We used search terms such as patient/consumer/user and visualization/graphic. Three reviewers systematically screened retrieved articles using prespecified eligibility criteria in Covidence (covidence.org), and independently assessed quality of included articles using the Mixed Methods Appraisal Tool. Characteristics of included articles and visualizations were extracted and synthesized following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Results

Of the 2,300 articles retrieved, 39 were included in this review. Most articles were excluded for including visualizations of non-health data (n=870) or visualizations intended for a clinicians or researchers (n=834). Sample sizes across the included studies varied drastically (369 ± 1,255). On average, participants were middle-aged (51 ± 11), non-Hispanic White (70%), female (60%), and had high education (77% college or higher). Few studies measured health literacy (n=5), numeracy (n=3), or graph literacy (n=2). Approximately half of the visualizations supported chronic condition monitoring and management while half were for general health maintenance (for example, lipid profiles). Visualizations included line graphs (35%), number lines (25%), bar graphs (16%), and icons (12%). Overall 77% of the visualizations used color; a traffic light color scheme was used most frequently. Many displayed longitudinal data (47%), 32% included legends, and 40% offered sentence-based interpretations of the data. Articles reported a range of methodologies and measures for assessing the visualizations, from qualitative explorations of patient attitudes and preferences to quantified measures of risk perception and comprehension.

Conclusion

Although findings suggest certain visualization types (number lines, icons) and components (color, contextual information) promote comprehension, the homogeneous patient samples, limited reporting of health literacy, numeracy, and graph literacy, and range of methodologies and outcome measures reported make assessment of optimal visualizations difficult. The informatics community and others engaged in this work would benefit from efforts to standardize best practices for the design, development, and evaluation of patient-facing visualizations.

This work was supported by NINR of the NIH under Award Number R00NR016275 (PI: Masterson Creber).

References

Similarity of Medical Concepts among Submitted Questions in Community-Driven Consumer Health Websites

Hamid Naderi, PhD1, Sina Madani, MD, PhD2, Behzad Kiani, PhD1, Kobra Etminani, PhD1
1Mashhad University of Medical Sciences, School of Medicine, Department of Medical Informatics, Mashhad, Iran; 2Vanderbilt University Medical Center, Department of HealthIT, Nashville, TN, USA

Introduction: Online Question and Answering (Q&A) communities are rapidly growing in World Wide Web. One of the major issues in these communities is duplicated questions, which result in a large number of analogous questions being posted on the websites. Hence, repeated questions make the search and browse for healthcare consumers quite frustrating. The ability to automatically categorize submitted questions based on structured topics on these websites and suggest similar Q&A to the users could potentially reduce the number of redundant questions. This study aims at comparing inter and intra-topic likeness amongst Q&As by incorporating concept-based similarity computing.

Methods: We randomly selected 3,000 Q&As from three popular online health communities including NetWellness, WebMD, and Yahoo Answers. Then, Unified Medical Language System (UMLS) concepts relevant to a set of selected questions from various topics were extracted by MetaMap (v2016) where the data source was limited to Consumer Health Vocabulary data source. Subsequently, Concept Unique identifiers (CUIs) were weighted by Term Frequency Inverse Document Frequency (TF-IDF) values. Finally, the cosine similarity among weighted vectors of UMLS concepts was computed (Figure 1). We also randomly selected 400 answered questions from NetWellness by 20 subject matter experts (20 Q&A for each expert) in order to compute expert-expert similarity scores. We generated profile for some experts based on answered questions. Then the similarity between profile of experts in same (expert-expert intra-topic similarity) and different topics (expert-expert inter-topic similarity) was computed.

Results: The average of intra-topic similarity in NetWellness, WebMD, and Yahoo Answers (0.095, 0.192, and 0.110 respectively) was about 10 times higher than the average of inter-topic similarity within the same communities for selected questions (0.012, 0.025, and 0.018 respectively). Similarity scores between the content of questions answered by subject matter experts in the same and different topics were calculated as 0.51 and 0.11, respectively.

Discussion: Comparing the results of intra-topic similarity and inter-topic similarity indicate the similarity between questions from same topics is significantly higher than the similarity between questions from different topics. It means the overlap between extracted concepts of questions from different topics is little. So using concept based method can classify questions of different topics automatically. Also Expert-expert intra-topic similarity shows there is considerable overlap between profiles of experts in the same topic. So concept based system can suggest users similar experts in a specific topic who have similar profiles. So concept-based similarity computing methods can be used in the development of intelligent Q&A retrieval systems with auto recommendation functionality for identification of comparable healthcare related questions as well as subject matter experts. Moreover, such methods can help community programmers define a process for automatic classification of questions and experts by topic, while integrating the content of several communities where healthcare consumers can find additional information from analogous questions that have been already asked and answered.

References:
Electronic Phenotyping of Patients with Heart Failure from a National Clinical Information Database

Masaharu Nakayama, MD, PhD¹,²; Chiaki Otomo, RN¹; Ryusuke Inoue, MD, PhD¹
¹Medical Informatics Center, Tohoku University Hospital, Sendai, Japan; ²Department of Medical Informatics, Tohoku University School of Medicine, Sendai, Japan

Introduction
A database of clinical information collected from several medical institutions, including national university hospitals and private hospital groups, and the Medical Information Database NETwork (MID-NET), have been available to the public in Japan since 2018 [1]. Tohoku University Hospital is one of the cooperative medical institutions of the MID-NET project. To collect clinical events known as electronic phenotyping, it is important to correctly extract data from clinical information, combine multiple pieces of information, and define the target disease. Herein, we examined a study to find patients with heart failure and to validate the accuracy of using MID-NET data.

Methods
A criterion to describe heart failure cases was determined according to clinical guidelines that were released by the Japanese Circulation Society, which is the largest community of cardiologists in Japan [2]. The data that we studied were based on the records from April 1 to December 31, 2013. The initial search conditions (initial rule) were based on disease names, examinations, and medications related to heart failure. We extracted and analyzed clinical data from MID-NET and found patients with heart failure. Two doctors, including a cardiologist, reviewed medical records and verified which cases were true. Precision and recall rates were calculated after the case assessments. Next, we examined a method to identify the elements to extract true cases correctly using machine learning with XGBoost in R (v3.4.1).

Results
There were 5,282 cases extracted by disease names that were related to heart failure. Of these, 2,799 cases corresponding to the initial rule were retrieved, and 200 cases were randomly sampled and assessed. There were 70 true cases. Thus, the precision rate was 0.350 and the recall rate was 0.912. A machine learning method using XGBoost revealed the correlation of heart failure with several factors, including the value of serum b-type natriuretic peptide (BNP), the linkage between the disease name start date and actual date of hospitalization, and medications for treatment of heart failure. Using this information, we were able to determine conditions contributing to improving the validity of the cases with heart failure. In this way, patients were extracted using the disease name as it related to heart failure and hospitalization within two weeks after the disease name starting date. Furthermore, the candidates were categorized into three groups according to the value of their serum BNP (High, Middle, and Low ranges). The high group was labeled “heart failure,” and the low group was excluded. In the middle group, candidates were additionally categorized according to their prescribed medication for heart failure. Our analysis indicated that the precision rate increased to 0.878, while the recall rate decreased to 0.697. The F-measure also increased from 0.506 to 0.777.

Conclusions
To find target cases from a large clinical database, precise electronic phenotyping is indispensable. A machine learning method can improve the accurate extraction of patients with heart failure. Leveraging large amounts of clinical data may be beneficial for making progress in medical research.

References
Automated classification of mobility activities in free text clinical narratives

Denis Newman-Griffis, MS,1,2 Ayah Zirikly, PhD,1 Pei-Shu Ho, PhD,1 Jonathan Camacho Maldonado, MD,1 Maryanne Sacco, MA,1 Alex Marr, BA,1 Albert M. Lai, PhD,3 Eric Fosler-Lussier, PhD2
1National Institutes of Health Clinical Center, Bethesda, MD; 2The Ohio State University, Columbus, OH; 3Washington University in St. Louis, St. Louis, MO

Introduction
Gathering and analyzing information on whole-person functional activities is an important part of assessing and improving an individual’s quality of life. However, reliably extracting information on activities from free text clinical records has proven challenging, in part due to a lack of coverage in controlled terminologies [1]. Recently, a dataset of 250 physical therapy records was annotated for descriptions of patient mobility [2], one category of whole-person activities defined in the International Classification of Functioning, Disability, and Health (ICF). Each of the 2,867 Action descriptions in the dataset was assigned one of thirteen 3-digit ICF codes indicating the type of mobility activity. The code names in the ICF match only a small subset of these descriptions, and a lack of integration with other controlled vocabularies limits their use for normalizing these codes. In this work, we formulated the normalization problem as multi-class classification, and experimented with a variety of popular classification models to understand what types of linguistic features can best distinguish different mobility activities.

Methods
We experimented with unigram text features and continuous word embeddings for classification, using both the words in the Action string and words within a fixed context window. We utilized several popular classification algorithms: Naïve Bayes (NB), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Deep Neural Networks (DNN). In addition, we also investigated the use of concept embeddings derived from averaging word embeddings for the definitions of each 3-digit code in the ICF. We compared these ICF code embeddings to the word embedding features derived from Action texts and contexts and chose the code with highest cosine similarity as the prediction. Finally, we used a deep neural network that took word embedding features of a specific context and ICF code embeddings as input and calculated a linear transformation of the code embedding for classification with the cosine similarity model.

Results

<table>
<thead>
<tr>
<th></th>
<th>NB</th>
<th>SVM</th>
<th>kNN</th>
<th>DNN</th>
<th>VecSim</th>
<th>VecSim+DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro Accuracy</td>
<td>29.7</td>
<td>53.3</td>
<td>38.4</td>
<td>37.3</td>
<td>52.4</td>
<td>82.6</td>
</tr>
<tr>
<td>Macro F1</td>
<td>16.8</td>
<td>44.1</td>
<td>17.7</td>
<td>30.0</td>
<td>42.8</td>
<td>51.6</td>
</tr>
</tbody>
</table>

Table 1. 10-fold cross-validation results, in %, for classifying Action mentions with 3-digit ICF codes. VecSim denotes vector similarity model; +DNN adds the context-sensitive linear transformation.

Overall micro accuracy and macro F1 (averaged over ICF codes) from our experiments are given in Table 1. Both the unmodified vector similarity model and the version with context-sensitive linear transformation outperform the other classifiers by a large margin, indicating that using embedding representations of the codes is more helpful for the normalization task than treating each as an orthogonal class label alone. In addition, using the context-sensitive transformation improves overall performance significantly ($p=0$, McNemar’s test). However, as shown by comparing the Micro Accuracy performance (weighted by sample) to Macro F1 (weighted by code), all methods are sensitive to skew in the dataset. Of the 2,867 total actions, 1,088 are labeled with the most frequent code (d450 – Walking), while the three least common codes each have fewer than 10 samples.

Conclusion
We found that neural concept embeddings achieved the best performance on classifying the type of mobility Actions, and that a context-sensitive projection of concept embeddings improved classification significantly. This indicates the importance of context-sensitive models for mobility information, and offers a potential direction for improving general clinical information extraction. Our results also indicate that robustly classifying rare codes is an important area of future work. This research was supported by the Intramural Research Program of the National Institutes of Health, Clinical Research Center and through an Inter-Agency Agreement with the US Social Security Administration.

References
Development and Assessment of Pharmacogenomic Clinical Decision Support for Thiopurine Methyltransferase Utilizing User-center Approach

Khoa A. Nguyen1,2, Himalaya Patel3, David A. Haggstrom1,3,4, Alan J. Zillich2, Thomas F. Imperiale1,3,4, Alissa L. Russ2

1Regenstrief Institute, Indianapolis, IN; 2College of Pharmacy, Purdue University, West Lafayette, IN; 3Health Services Research and Development Service (CIN 13-416), Richard L. Roudebush VA Medical Center, Indianapolis, IN; 4Department of Medicine, Indiana University School of Medicine, Indianapolis, IN;

Introduction
Poor design of Pharmacogenomics clinical decision support tools (PGx-CDS) can introduce errors and create alert fatigue instead of improve treatment efficacy and safety for patients1,2. The objective of this study was to develop and evaluate a prototype of a PGx-CDS interface with user-centered design methods.

Methods
This study had two phases: In phase I, we conducted qualitative interviews to assess providers’ information needs from two healthcare systems. Interviews were audio recorded and transcribed for analysis through a combination of inductive and deductive qualitative analysis to develop design requirements for a PGx-CDS user interface. In Phase II, we conduct usability evaluates of the prototype in a simulated examination room. Physicians, specializing in oncology and gastroenterology, completed two fictitious patient scenarios involving PGx-CDS for thiopurine methyltransferase (TPMT). Qualitative data were collected via Morae® video and audio recording using a “think aloud” technique and debriefing interviews. Quantitative usability data were measured on four established measures: learnability, usability errors, efficiency, and satisfaction utilizing the 19-item Post-Study System Usability Questionnaire.3

Results
Fourteen providers participated: 10 were interviewed in phase I and 7 provider completed usability testing in phase II (3 providers participated in both phases). Most (90%) providers were interested in PGx-CDS systems to help improve medication efficacy and safety. Interviews yielded 8 themes (Table 1) sorted into two main categories: health care provider’s views on PGx-CDS and 2) important design features for PGx-CDS. Usability testing showed high provider satisfaction from the prototype (Table 2).

Table 1. Main themes identified from interviews

<table>
<thead>
<tr>
<th>Healthcare providers’ views on PGx-CDS</th>
<th>Important design features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The need for PGx-CDS for TPMT</td>
<td>5. PGx-CDS content and display</td>
</tr>
<tr>
<td>2. Impact of PGx-CDS on clinical workflow</td>
<td>6. Genetic result content and display</td>
</tr>
<tr>
<td>3. Lab testing preferences</td>
<td>7. References within PGx-CDS</td>
</tr>
<tr>
<td>4. Perceived barriers to PGx implementation</td>
<td>8. PGx care coordination</td>
</tr>
</tbody>
</table>

Table 2. Satisfaction score from 19-items Computer System Usability Questionnaire

<table>
<thead>
<tr>
<th></th>
<th>System usefulness</th>
<th>Information quality</th>
<th>Interface quality</th>
<th>Overall satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>GI specialist Median (range)</td>
<td>7.0 (6.0-7.0)</td>
<td>7.0 (4.8-7.0)</td>
<td>6.0 (4.7-7.0)</td>
<td>6.0 (5.0-7.0)</td>
</tr>
<tr>
<td>Oncologist Median (range)</td>
<td>6.5 (6.0-7.0)</td>
<td>6.5 (6.0-7.0)</td>
<td>6.5</td>
<td>6.5 (6.0-7.0)</td>
</tr>
</tbody>
</table>

Conclusion
This study used a user-centered approach to develop a PGx-CDS for TPMT. The results provides guidance for the development of a PGx-CDS, particularly for biomarkers such as TPMT

References
Longitudinal Satisfaction with Information for Breast Reconstruction Patients

Krista M. Nicklaus, MSE1,2, Mary Catherine Bordes, BS2, Jun Liu, PhD2, Deepi Chopra, MBBS, MPH2, Greg P. Reece, MD2, Summer E. Hanson, MD, PhD2, Mia K. Markey, PhD1,2

1The University of Texas at Austin, Austin, TX; 2The University of Texas MD Anderson Cancer Center, Houston, TX

Introduction

Prior studies have demonstrated psychosocial and quality of life benefits for patients who undergo breast reconstruction after mastectomy, but patients’ decision-making about breast reconstruction is challenging, especially with respect to the available preoperative information1. The goal of this study is to investigate patients’ satisfaction with the information they received from their healthcare providers over multiple time points after their initial reconstruction surgery to identify unmet information needs and factors associated with information gaps.

Materials and Methods

We have previously collected health data, image data, and psychosocial measures from 505 breast reconstruction patients from The University of Texas MD Anderson Cancer Center from 2011-2014 (funded by NIH R01CA143190 and R01CA203984). This study used the BREAST-Q v.1 Post Reconstruction Satisfaction with Information module, which includes 15 items to rate satisfaction with information provided by the plastic surgeon about issues such as healing time, complications, and new breast look and feel. This module was administered at four time points after patients’ initial or definitive reconstruction surgery.

For initial analysis, we selected 118 patients who had satisfaction with information scores at the first two time points and complete demographic data, including assessments of preoperative body image investment (ASI-R) and body image concerns (BIS). We used cluster analysis to look for trends in how satisfaction with information changed over time. Figure 1 demonstrates the results of k-means clustering based on the satisfaction with information scores at T1 (3 months postoperative) and T2 (12 months postoperative) as well as the difference between the scores at T1 and T2.

Results and Conclusion

Cluster analysis of longitudinal satisfaction with information revealed patterns that are not simply a reflection of differences in preoperative body image investment or body image concerns. For example, women in cluster #1 and cluster #2 had similar ASI-R and BIS scores, but different satisfaction patterns over time. The clusters are being further investigated with the aim of identifying preoperative characteristics that predict which patients are mostly likely to benefit from postoperative information interventions, such as just-in-time delivery of information about surgical complications. This study represents the use of patient-reported outcomes and psychosocial data to improve healthcare information delivery to breast reconstruction patients.

References

Using Electronic Health Record Data to Identify Patients with Prediabetes

Thomas J. O’Byrne, MAS1, M. Regina Castro, MD2, Che G. Ngufor, PhD1, Gyorgy Simon, PhD3, Pedro J. Caraballo, MD2

1Department of Health Science Research and 2Department of Medicine Mayo Clinic, Rochester, MN; 3Department of Medicine, University of Minnesota, Minneapolis, MN

Introduction

It is estimated that 30% of the population suffers from pre-diabetes, in which glucose levels are above normal but still below current criteria for diabetes. Pre-diabetes represents a state of increased health risk and is an important risk factor for the development of overt diabetes. Current evidence shows poor recognition and limited management in routine clinical practice. There are challenges to identify patients with prediabetes including lack of knowledge, confusing glucose levels, lack of documentation in the electronic medical record, etc. Current national clinical guidelines recommend screening and management of patients with prediabetes as the most cost-effective intervention to confront the growing diabetes epidemic. Our aim is to define algorithms to accurately identify patients with pre-diabetes by using available electronic health record (EHR) data obtained during routine care. Herein, we present our initial findings related to feasibility, validation of the algorithm and clinical categorization.

Methodology

The study included 48,038 adult (≥ 18 y-o), non-pregnant residents of Olmsted County, Minnesota, who received primary care at Mayo Clinic between 1999 and 2014. We used electronic searches to gather data and used “fasting” glucose, A1c, diagnoses and medications related to diabetes in an algorithm to allocate all subjects into 4 mutually exclusive and exhaustive categories: “Diabetes”, “Pre-diabetes”, “Normoglycemia”, and “No-data” (no available data applicable to the other categories). Validation of the algorithm was performed using a powered sample of patients (n=159) with blind and manual clinical review of the comprehensive EHR by an expert clinician. To evaluate agreement, we used Kappa, Weighted Kappa and secondary measures of agreement. Two sensitivity analyses were performed to examine the ability of the algorithm to correctly categorize patients. The first sensitivity analysis used only 1 indicator of diabetes disease instead of 2, as used in the main analysis, and the second sensitivity analysis used 3 indicators.

Results

The proportion of subjects allocated to the 4 categories in the large cohort, Diabetes, Pre-diabetes, Normoglycemia and No-data were 9.5%, 29.6%, 28.5% and 32.4%, respectively. The random sample had similar characteristics to the rest of the cohort except for having lower mortality. There was a strong agreement between the algorithm and the manual clinical review, Kappa 0.85 (95%CI 0.78-0.92, p<0.0001) and Weighted Kappa 0.90 (95%CI 0.84-0.96, p<0.0001). The results of the sensitivity analyses are shown in the Figure. The proportions of No-data and Normoglycemia were similar between the manual clinical review and both sensitivity analyses. However, proportions of Diabetes and Pre-Diabetes were different. The changes made to the algorithm for the sensitivity analyses affected classification of patients only into the Diabetes and Pre-Diabetes categories.

Conclusion

Our data showed that it is feasible to identify clinically helpful phenotypes describing the different levels of dysglycemia by using EHR data obtained during routine clinical care. The use of discrete clinical categories should facilitate implementation of clinical decision support and longitudinal plan of care, using tools currently available in most widely used EHR. These interventions would support current best strategies to minimize the impact of the diabetes epidemic and associated devastating complications.
Building a LVEF Extraction Pipeline for SOAP Notes
Adetomiwa Oguntuga, MS¹, Jordan Overcash MS¹, Nam Nguyen, MS¹
¹Veradigm, San Francisco, CA

Introduction
Congestive Heart Failure is a common chronic medical condition that affects the pumping power of muscles in the heart. Left Ventricular Ejection Fraction (LVEF) is a quality measurement recommended by the American College of Cardiology Foundation, the American Heart Association and the Physician Consortium for Performance Improvement, to access pumping power of the left ventricular wall of the heart.

We build a prototype information extraction pipeline in python, using natural language processing (NLP) techniques to identify and extract these assessments from SOAP notes on Practice Fusion EHR data.

Methods
500 SOAP notes from Practice Fusion’s EHR database were identified as possibly containing LVEF scores, using keyword search. The keywords used were: left ventricular ejection fraction, lv ejection fraction, ejection fraction, e.f, ef, heart failure reduced ejection fraction, hhref, heart failure preserved ejection fraction and hhpief.

These notes were run individually through a prototype extraction pipeline, coded in Python 3 (version 3.6)².

This pipeline used functionality from the spacy NLP library (version 2.0)³ to chunk notes into a list of sentences. Sentences containing LVEF assessments were chunked further based on the presence of commas or semi colons into smaller chunks. Commas present in dates are ignored and the dates are converted into a numerical MM/DD/YYYY or MM/YYYY format. The final list of chunks is then processed by using regular expressions, to extract the LVEF assessment and the date associated with the assessment (if present in the note).

LVEF assessments come in three types: a percentage (e.g. 45%), a range of percentages (e.g. 40-50%), or a relative percentage (e.g. >40%). Precision, recall and F1 score were calculated by assessment type.

Results
The results presented in the table below are from sentences identified as containing an LVEF assessment (Table 1).

<table>
<thead>
<tr>
<th>Assessment Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>Assessment as Percentage</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>Assessment as Range</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>Assessment as Relative #</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessment Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.93</td>
<td>0.98</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 1: Precision, Recall and F1 Measure Score for Prototype Pipeline

Discussion and Conclusion
Our overall goal is to build a tool that can extract LVEF assessments from our unstructured data sources to enrich our structured data to better enhance our life science analysis studies. We have been able to extract LVEF scores and dates from identified sentences with a high degree of accuracy. Though some data had an associated date, the majority of extracted LVEF scores did not. Future work will focus tying the extracted LVEF score to an encounter. In addition, we need to access the recall and precision of the logic that identifies sentences with LVEF.

References
Determinants of Medication Adherence in Sickle Cell Disease Using the World Health Organization Model

Kinsley E. Ojukwu1, Christopher Simpson1, Amol Utrakar, MD2, Whitney Allen3, Laurie Novak, PhD, MHSA1, Robert M. Cronin, MD1
1 Vanderbilt University, Nashville, TN, USA, 2 Columbia University, New York, NY, USA, 3 Meharry Medical College, Nashville, TN, USA

Introduction
Over 100,000 individuals in the United States are affected with sickle cell disease (SCD)1, and the disease has great prevalence among those with healthcare and healthcare information technology disparities. Hydroxyurea is a medication that decreases morbidity and mortality in SCD, yet has poor adherence2. In a 2003 report on adherence to long term medical therapies, the World Health Organization (WHO) described a model that organized determinants of medication adherence into five interacting dimensions3. There are limited data exploring both barriers and promoters of medication adherence in SCD, and none analyzed these determinants of medication adherence in the context of the WHO model.

Methods
A mixed methods approach was used for this study. Participants were selected via a convenience sampling of patients seen at outpatient clinics affiliated with the Vanderbilt-Meharry Center for Excellence in SCD. Demographic data was collected through the use of surveys. Semi-structured interviews were conducted between September 2016 – July 2017. Interviews were audio recorded, transcribed, and analyzed using Dedoose Version 8.1.8. Analysis was done with an inductive-deductive approach4. Open coding methodology was utilized, and a codebook developed to identify promoters and barriers of medication adherence. These codes were subsequently categorized based on the WHO Model. Codes were reviewed by two independent coders, and categorization done via consensus of three authors (KEO, CS, and RMC).

Results
Forty-six participants ranging from age 14-61 were enrolled. All participants carried a diagnosis of SCD. The median age was 28.5, and a majority were female (n=24). Most self-identified as Black or African American (n=42). Fifty distinct codes were applied with 38 barriers, and 14 promoters. Twenty-six were patient-related factors (e.g., fear of addiction – barrier), 11 social & economic (e.g. theft of medications – barrier), six condition related (e.g. subjective feeling of well-being – promoter), five therapy related (e.g. dosing schedule – barrier), and four health system (e.g., a phone application to track medications or appointments – promoter) (Table 1). Some codes fit multiple categories, (e.g., sense of responsibility) – a barrier in younger ages, but a promoter in older ages, and stigma, both a barrier and promoter, and categorized into both social/economic and patient related categories.

Conclusion
This study found that although a majority of determinants were patient-related in nature, about half came from other dimensions, showing the multi-dimensional nature of medication adherence in SCD. We also discovered unique determinants in this population, such as the barrier of medication theft. Limitations to this study include being performed at a single center with a convenience sampling method of individuals from clinic, which may lead to lack of generalizability and selection bias respectively. However, many determinants extracted from our research are concordant with determinants previously described in literature, such as forgetfulness and concerns about medication side effects. Using the WHO framework to understand medication adherence determinants in SCD, a disease affecting those with healthcare and healthcare information technology disparities, can enable the creation of informatics solutions that can help increase medication adherence, and potentially help bridge the digital divide.

References

| Table 1: Select excerpts from semi-structured interviews |
|---------------------------------|---|---|
| **Dimension** | **Code** | **Quote** |
| Social & Economic Factors | Theft | “… but you got to risk someone stealing the medication…They can see prescription bottles and they don’t have time to read them so they stick everything you know…” |
| Patient Related Factors | Responsibility | I just have to take the meds… I just have to do what I can do while I’m here…I have a good life. I have a nice house, nice car, I work hard… That’s why I take my meds everyday. I got a lot to live for. |
| Patient Related Factors | Stigma | “If you have a certain time you’re supposed to take it then that’s probably during school then some people might ask, “Why you have to take this?” … Then you kind of like not want to explain or just don’t want everybody to know…” |
Untapped Potential of Clinical Text for Opioid Surveillance

Amy L. Olex, MS, Tamas Gal, PhD, Majid Afshar, MD, MSCR, Dmitriy Dligach, PhD, Niranjan Karnik, MD, PhD, Travis Oakes, BS, Brihat Sharma, BS, Meng Xie, MS, Bridget T. McInnes, PhD, Julian Solway, MD, Abel Kho, MD, William C. Cramer, MS, F. Gerard Moeller, MD

Virginia Commonwealth University, Richmond, VA; Loyola University Medical Center, Chicago, IL; Loyola University, Chicago, IL; Rush University, Chicago, IL; University of Chicago, Chicago, IL; Northwestern Medicine, Chicago, IL; Virginia Commonwealth University Health System, Richmond, VA

Introduction

Accurate surveillance is needed to combat the growing opioid epidemic for effective resource mobilization. Current methods primarily rely on diagnostic codes, which are specific, but not sensitive, potentially resulting in many opioid overdose encounters being missed by current surveillance methods. Recent application of Natural Language Processing (NLP) to clinical notes indicates there is a wealth of untapped, opioid-specific information that could be used to improve surveillance. To investigate the potential volume of missed opioid overdoses, we compare the number of overdose encounters identified by diagnostic codes and a rule-based NLP pipeline from two different medical systems.

Methods

Emergency department (ED) opioid overdose (OOD) encounters between 9/1/16 and 9/1/17 for patients 18 years or older were identified using two methods: 1) claims data using ICD-10-CM codes and 2) ED notes using NLP. Encounters identified by claims data were associated with one or more overdose-related ICD-10-CM codes. NLP processed the first 24-hours of ED notes using a rule-based approach. Briefly, we utilized CLAMP v1.5.0 to construct the NLP pipeline, which includes standard pre-processing NLP modules, a custom dictionary utilized by the Named Entity Recognition module to identify terms associated with OOD, and a section identification module. Pre-processed notes were subjected to rules that included removing all annotations from patient education sections, and identifying the patients response to Narcan. Virginia Commonwealth University (VCU) and Loyola University (LU) ran both methods for a cross-institutional evaluation using the same NLP pipeline and ICD-10-CM codes.

Results and Discussion

At both medical institutions, the NLP pipeline was able to identify more than 60% of the encounters found using ICD-10-CM codes plus additional encounters not identified by the claims data. At VCU, 1,228 encounters were classified as OOD with claims data identifying 566 versus NLP’s 1,011 (349 were found by both methods). At LU, 2,584 encounters were classified as OOD with claims data identifying 84 and NLP 2,556 (56 were found by both methods). The additional encounters identified by NLP had ICD-10-CM codes related to altered mental status and opioid abuse/dependence, indicating that NLP is identifying OOD-related encounters missed by claims data. Random sampling of 30 encounters at VCU revealed about 45% identified by NLP were false positives. These included chronic pain patients and patients with poor mental status from other medical conditions who received Narcan as empiric treatment prior to definitive diagnosis.

Conclusion and Future Work

Our analysis showed that the NLP pipeline was able to successfully capture additional OOD encounters missed by claims data. Refinement of the NLP pipeline by implementing additional rules that do not rely on Narcan administration and ignoring all annotations in the prescription section should aid in eliminating false positives. Additionally, the completion of an annotated gold-standard data set will provide concrete performance metrics and highlight additional areas in need of improvement. Incorporating sophisticated NLP and machine learning techniques into current diagnostic code-based methods has the potential to improve surveillance and obtain more accurate information on the incidence of opioid overdoses.

†SAMHSA opioid-related death specifications minus “adverse effect” codes and adding T40.69 and T40.60.
Evaluating Interactive Clustering for Biomedical Information Retrieval

Michael Segundo Ortiz¹, Kazuhiro Seki, Ph.D.², Javed Mostafa, Ph.D.¹
¹University of North Carolina, Chapel Hill, NC;
²Konan University, Kobe, Hyogo, Japan

Problem

Large biomedical research databases such as PubMed rely on keyword-based queries and ranked-retrieval methods. This method is an emerging barrier in advancing research due to the large body of publications, even within specialized sub-fields. For instance, a single search in PubMed for “breast cancer” returns a long list of more than 380,000 articles. It is a non-trivial task for researchers to go through a ranked list and digest content, let alone establish logical connections within the content. A possible solution is to employ a system with a cluster-based browsing interface which potentially enables more effective concept and document retrieval.

Objectives

Our objectives for this work were to carefully create document collections (data sets) by considering underlying topic groups manually assigned by human experts (MeSH), and identify essential components and methods for discovering quality document clusters in order to enable the production and use of a dynamic cluster-based browsing system.

Methods

We retrieved PubMed articles related to breast cancer with the query "breast neoplasms" on September 21, 2018. It should be clarified that the MeSH labels were only used to create the data sets and then to evaluate downstream clustering. The methodology of clustering itself is unsupervised, therefore class labels were removed for the experiments. We hypothesized that full-text data may offer more lexical diversity and, consequently, may result in higher quality clusters. For document representation, we used a vector space model. Additionally, we explored the effectiveness of dimensionality reduction, specifically latent semantic analysis (LSA). For cluster analysis we used the maximin (maximum-minimum), k-means++, and spectral clustering algorithms. As an internal criterion, we adopted the silhouette coefficient. For external criteria, purity and adjusted mutual information were used.

Evaluation

We observed that k-means++ combined with keyword discovery methods and LSA worked the best compared with other parameter configurations. Contrary to our expectations, full-text data did not help increase the cluster quality despite that it required around 13–20 times more processing time. Also, the increased lexical diversity of full text made the clustering process more complex and the results less precise. Overall, using abstracts appeared to achieve a better balance between lexical diversity, cluster quality, and time complexity.

Conclusion

Although our work sheds light on the fundamental components of evaluating cluster quality for biomedical information systems, optimum parameter values and combinations depend on input data, which dynamically change in the cluster-based browsing paradigm. There is a need for a mechanism to automatically and intelligently adjust parameters for highly variable inputs to achieve externally valid results.

Future Plans

We are currently building a prototype biomedical information system for dynamic cluster-based browsing based on our findings. Follow [http://vzlib.unc.edu/dcb2/](http://vzlib.unc.edu/dcb2/) to explore the prototype.

Acknowledgments

The NIH-NLM T15 grant 5T15LM012500-02, United Health Foundation’s ENABLE grant, and MEXT, Japan, provided support for this work.
Investigating female–male differences in risk factors for myocardial infarction using OHDSI tools

Anna Ostropolets, Linying Zhang, Jami Jackson Mulgrave, George Hripcsak
Department of Biomedical Informatics, Columbia University, New York, USA

Background
Coronary heart disease (CHD) has been the leading cause of mortality and disability across the world. There is no consensus regarding the sex-specific influence of additional risk factors on outcomes of CHD: while Millett et al. and Anand et al. reported higher risk of myocardial infarction associated with diabetes mellitus (DM) type II in women compared to men, Shah et al. found no significant sex-specific difference. Similar disagreement has been observed for obesity, smoking and atrial fibrillation (AF). We tested the assertion that given a complex study result from the literature, we could quickly reproduce the study, mimicking the variables by generating matching phenotype definitions using standard vocabularies, and carrying out the statistical analyses on electronic health record data. We choose Millett’s study for its nested hypothesis that compares the relative effects in two groups.

Methods
We performed our analysis on the Columbia University Medical Center database that contains de-identified electronic health records of 6.3 million patients stored in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). We studied patients aged at least 30 years old as of 01 January 2012 (index date) who had never experienced major cardiovascular events (stroke, transient ischemic attack, myocardial infarction (MI), episodes of angina) prior to index date. To study the impact of risk factors, we leveraged the OMOP Standardized Vocabularies to obtain the comprehensive list of SNOMED diagnosis codes for hypertension, AF, obesity, DM types I and II, and smoking status. Additionally, for obesity we used LOINC codes for measurement of body mass index of 30 kg/m2 or higher. We used Observational Health Data Sciences and Informatics (OHDSI) Atlas web interface to generate phenotype definitions. We used a Cox proportional hazard model adjusted for age, hypertension, DM, obesity, and smoking status to estimate the baseline hazard ratio for MI comparing to men. To determine the hazard ratios for the risk factors, we created interaction terms between the risk factor and sex. Like Millet et al, we adjusted diabetes and hypertension for each other and for smoking status; obesity and use of lipid-lowering medications; AF for DM, hypertension, obesity, smoking status, and use of lipid-lowering medications; and obesity for smoking status. Additionally, we studied the effects of risk factors in different age groups by splitting patients into three age groups (<50, 50–59 and >60 years old). To obtain the ratio of hazard ratios for women comparing to men we used the interaction terms between these risk factors and sex.

Results
We identified 3,076,220 eligible patients (57.6% women and 42.6% men), with the mean age at index date 52±14.1 and 54±14.0 respectively. Baseline prevalence of diabetes, hypertension, obesity, and reported smoking in men and women did not differ significantly. Of 17661 incidents of MI that were recorded, the majority belonged to men (64%). We found that women had a lower overall risk of MI than men (women-to-men ratio of HR 0.45, 95% CI 0.44 – 0.46) after adjusting for the risk factors and age. This trend attenuated with increasing age (ratio of HR 0.53, 95% CI 0.48 – 0.59). Smoking was associated with an increased risk of MI in both women (HR 2.94, 95% CI 2.54 – 3.39) and men (HR 2.23, 95% CI 3.02 - 4.42); the difference between men and women was not statistically significant. We observed the same pattern for obesity: obese women developed a higher risk of MI (HR 4.62, 95% CI 4.11 – 5.2) compared to men (HR 4.02, 95% CI 3.55-4.57), although the women-to-men ratio was not significant. Hypertension, DM type I and type II and AF were also associated with higher attributable risk in women than men, with a women-to-men ratio ranging from 1.5 (DM type I) to 1.88 (AF). Presence of DM type II, hypertension, or AF was associated with a higher risk of MI in both sexes: hazard ratio varied from 2.0 (DM type II) to 2.19 (AF) in women and from 1.16 (AF) to 1.34 (hypertension) in men.

Conclusion
We were able to duplicate the study on electronic health record data stored in OMOP CDM. The phenotype definitions were based on logical and temporal constructs applied over classes from the OMOP Standardized Vocabularies using OHDSI’s Atlas web interface. Overall, we found tighter confidence intervals that overlapped the original intervals. From our experience, OHDSI Atlas facilitated phenotype generation and reduced the effort on initial data analysis. On the other hand, we could not fully implement the statistical model in Atlas, so it took several iterations to parallel the study’s statistical analysis in R. Therefore, the lesson for OHDSI is that the phenotyping component of Atlas is sufficient, but the statistical analysis component must be expanded.

References

1695
Developing an Ontology for Daily-Living Routines of Patients with Chronic Conditions

Mustafa Ozkaynak, PhD

1University of Colorado | Anschutz Medical Campus, Aurora, CO

Introduction

Individualized therapy plans for chronic disease management require accounting for the daily-living routine of patients. Daily-living routines can be an effective operationalization of social determinants of health, which play an important role in health outcomes. If these routines can be captured and integrated into electronic health records (EHR), embedded decision support tools can utilize data from these routines (along with other patient information e.g. laboratory results, co-morbidities), to support clinicians develop individualized treatment plans. However, integration of these daily routines into the EHR, first requires their representation in a machine interpretable way. Such representation can be accomplished using ontologies. An ontology facilitates development of individual instances of well-described classes or concepts that constitute a knowledge base. The purpose of this study was to develop an ontology for daily-living routines of patients with a chronic condition requiring anticoagulation therapy.

Methods

Protégé1 was used to develop an ontology based on qualitative data collected to examine health management of anticoagulation therapy. Data was obtained from 39 patients from a single clinic through interviews and journaling. The first step of ontology development was conceptualizing daily-routines as workflows. The temporal features of the concept of workflow is beneficial to characterize daily-routines. The second step was a middle-out strategy for ontology development i.e., starting from salient and moving to general and specific concepts and relationships. We defined classes, properties of classes (slots), and subslots by qualitatively analyzing data and conducting a review on workflow in health informatics, human factors, and ergonomics literature. Previously developed ontologies (i.e. HELIS2 and SPHERE3) informed the specification, knowledge acquisition, conceptualization and integration of the ontology.

Results

We developed a preliminary ontology that represents daily-living routines and health management activities within these routines for patients on anticoagulation therapy. The ontology included: (1) activities, roles, use of information, and use of tools; (2) temporal organization of activities and rules in the form of constraints; and (3) physical, organizational, social, and cultural contexts. To evaluate the flexibility of the ontology, we developed instances for 16 of 39 participants with different characteristics in terms of age, gender, annual income and duration of therapy. The ontology is based on a corpus of qualitative data, as opposed to an expert panel, is the more common approach.

Conclusion

Future EHRs will have to include extra-clinical health-related workflow information4. The developed ontology allows production of machine-interpretable knowledge and makes possible clinical decision support that is sensitive to the patient’s daily-living routine and contexts. The therapy plans that are assisted by such decision support, can yield better adherence, hence improved patient outcomes. Future work includes validating the ontology by exploring the application of this ontology in patients with other chronic conditions. This study was supported by AHRQ under award number R03HS024092 (PI: Ozkaynak).

References


1696
Development and Usability Evaluation of a Facebook-based Program for Pediatric Cancer Patients

Bu Kyung Park, PhD, RN
1Kyungpook National University, College of Nursing, Daegu, Republic of Korea

Abstract
The aim of this study was to develop and evaluate the usability of a Facebook-based program (“Healthy Teens”) for pediatric cancer patients. The process of the program’s development and evaluation was 1) focus group interview for needs assessment with 12 pediatric cancer patients and 12 of their family members, 2) content development and expert review, 3) usability evaluation with 11 pediatric cancer patients. The program was perceived by pediatric cancer patients to be usable and useful in improving their health-related interventions.

Introduction
Social networking sites such as Facebook have become a popular venue for teenagers. Concurrently, the prevalence of smartphones among youth and the expansive reach of the Internet have enabled various new educational methods. With the increasing survival rate of pediatric cancer patients, there is a greater need for education about self-management at home after discharge.

Objective
The aim of this study was to develop and evaluate the usability of a Facebook-based educational program (“Healthy Teens”) for pediatric cancer patients.

Methods
Program development and evaluation progressed as follows: 1) focus group interview for needs assessment, 2) content development and expert review, and 3) usability evaluation. First, we conducted a focus group interview with 12 pediatric cancer patients and 12 family members. Qualitative content analysis using the inductive coding approach was performed on the interview transcripts. Second, an expert panel developed the program content, based on social cognitive theory.1 Then, three experts in the field of pediatric cancer reviewed the program content. Lastly, we conducted a usability evaluation of the Facebook-based program with 11 pediatric cancer patients. We improved the program’s usability through iterative cycles of evaluation and adaptation. We measured participants’ perceived usability with the Perceived Health Web Site Usability Questionnaire (PHWSUQ)2 and their eHealth literacy with the Korean eHealth Literacy Scale (K-eHEALS)3.

Results
Based on the results of the focus group interview and expert review, we finalized the Healthy Teens program to contain eight modules comprising multimedia, printable materials, quizzes, and discussion topics.

Results of the usability evaluation indicated that participants (age range: 13 – 20 years old) deemed the appearance and content of the program appropriate, and the program itself, easy to use. Most participants recommended to add more graphics and explanations on medical terms used for their treatment. Participants’ perceived usability of the program was 60.09 (range: 12-84), and their eHealth literacy was 27.55 (range: 8-40).

Conclusion
Overall evaluation of the Healthy Teens program on Facebook revealed that pediatric cancer patients perceived the program to be usable and useful in improving their health-related interventions for self-management at home.

References

1697
A Case Study of Remote Monitoring Processes in Texas

Sulki Park, MS¹,², Hye-Chung Kum, PhD¹,², Mark Lawley, PhD²
¹Health Policy & Management, Texas A&M School of Public Health, College Station, TX
²Industrial and Systems Engineering, Texas A&M University, College Station, TX

Introduction

Remote monitoring (RM) is likely to be implemented more widely as policy makers break down legal and regulatory barriers, as well as providers focus on improving RM services to meet the needs of patients and caregivers. Medicare expands reimbursement for RM releasing new procedure codes in 2019. Some states pursue their own unique set of reimbursement policies in their Medicaid programs. Texas Legislature has passed several bills since 2013 to establish RM for patients with hypertension or diabetes as a reimbursable health service under Texas Medicaid. Designing cost-effective RM processes, for both the automated computer processes and human processes, is the key to a successful implementation.

Objective

The purpose of this research was to evaluate RM processes by conducting a case study of one RM company in Texas. This study will provide further insight into understanding the key elements of the RM processes.

Methods

We conducted a population-based descriptive analysis of the processes for RM using a company data from 2016-2018. Only hypertensive or diabetic patients, enrolled in Medicaid, were eligible for this study. Blood pressure (BP), blood glucose (BG), or both were monitored daily after enrollment. Automated alerts were generated when 1) physiologic parameters were not transmitted on time (i.e., adherence alerts), or 2) the vital sign was out of acceptable range (i.e., clinical alerts), set by the physician. Both alerts were followed up with a phone call to the patient or caregiver. Non-clinical staff followed up on adherence alerts to trouble shoot in order to receive the daily reading. Clinical staff followed up and classified the clinical alerts as red, yellow, or green, depending on their severity and followed set protocol (e.g., red and yellow means, follow-up should occur both over phone and email to the clinical contact, and green should follow up over email). Weekly reports were transmitted to primary physicians with the vital signs and alert classification. In our study, daily records and alerts were analyzed.

Results

A total of 2,075 patients got monitored daily for BP, BG, or both from 2016-2018 by 66 physicians. The average age of the patients was 72 (SD ± 12), with 66% being female and 78% living in the South Texas area. On average, patients were authorized for monitoring for 363 days (SD ± 240). Of the authorized days, on average, 46% of these days (167 days ((SD ± 151)) required non-clinical follow-up to troubleshoot (e.g., battery issues, transmission issues) or remind patients take a reading. With these follow-up efforts, on average, 76% of the days (274 days (SD ± 201)) actually had a vital sign sent to the company for monitoring. Of the patients being monitored for BP, on average, 39% of monitored days (107 days (SD ± 107)) got clinical alerts, having a record that is out of acceptable range, and of BG patients, on average, 6% of the days (16 days (SD ± 35)) got clinical alerts. 93% of BP patients set diastolic/systolic BP as 60/90-90/160 mmHg to be an acceptable range and 89% of BG patients set 60-350 mg/dL to be acceptable range. On average, 87% of BP clinical alerts (93 days (SD ± 96)) and 75% of BG clinical alerts (12 days (SD ± 28)) were classified as being green, which means no clinical intervention needed, whereas red or yellow require clinical intervention.

Discussion and Conclusion

On average, about half of the authorized days required follow-up to maintain daily monitoring. These non-clinical calls with the patients constitute the largest time interacting with patients at the company. Even with these calls, on average, patients missed 2.4 of 10 days transmitting their vital sign. In addition, most automatic clinical alerts were classified as green that required no clinical intervention pointing to the difficulty of relying only on daily vital signs to detect clinically meaningful events. Thus, to establish an effective RM platform, it will be important to provide for non-physician clinical expertise to interact with the patients to differentiate between automatic alerts that need physician intervention and those that do not as well as non-clinical time to troubleshoot and maintain daily monitoring.
Extracting and Standardizing Social Determinants of Health Diagnosis from Problems List from a Large Ambulatory Electronic Health Record (EHR) Database

Svati Patel, MHS MS
Nam Nguyen, MS
Allscripts, Chicago, IL
Practice Fusion, San Francisco, CA

Introduction
There is increased recognition that social determinants of health (SDH) outside and within the traditional healthcare setting play a powerful role in shaping patient health outcomes (1,2,3). We developed a methodology to extract and analyze both structured and unstructured SDH diagnosis from problems list data from a large ambulatory EHR database.

Methods
Problems list data from an Allscripts EHR database containing deidentified health information for nearly 34 million unique individuals was mined for SDH diagnoses recorded in the ambulatory setting. 24% of the problem list diagnostic data was recorded as un-coded, unstructured free text. ICD-10, ICD-9, and SNOMED codes were used to extract SDH from the structured problem list data. SDH was extracted from unstructured data using a pattern matching search algorithm that utilized keywords, word pairs, and exclusion words. We also developed a mapping algorithm using the SDH framework found in ICD-10 codes Z55-Z65 to standardize captured data.

Results
Extraction methods yielded 1.45 million unique SDH diagnoses recorded within the Allscripts EHR database. These diagnoses were obtained using 96 icd-10 codes, 45 icd-9 codes, 619 SNOMED codes, and 1,008 unstructured SDH terms. 77% of the diagnoses were recorded as structured data; 23% of were recorded as unstructured data. 90% of the diagnoses found within the Allscripts EHR database were mapped to 14 (out of 81) SDH diagnoses. The top three SDH recorded were issues associated with divorce or separation, death or disappearance of a family member, and unemployment. Further, 90% of the SDH diagnostic data was recorded by 87 unique health systems (31%).

Discussion
Analysis reveals that SDH information can be recorded within a patient’s list of medical diagnoses, the problem list. However, the vast majority of the diagnostic SDH data is being recorded by only a small percentage of health systems represented in the Allscripts EHR database and regarding only a small percentage of SDH topic areas. Extraction of this data requires use of techniques that identify both structured and unstructured methodologies.

Conclusion
Though the problem list yields mineable SDH data, a full exploration of other EHR data is needed to understand the amount of SDH data available from ambulatory EHRs. Other possible sources of SDH data from ambulatory EHR databases that warrant exploring include SOAP notes and chief complaints. Future work will improve the recall and precision of the extraction from problem list data, adapt the extraction methodology to work with SOAP notes and chief complaint data, and access the full amount of SDH data that is currently recorded in ambulatory EHR data.

Acknowledgements
Fatima Paruk, MD and Dan Connors, PhD.

References
A Dataset Recommendation System for Researchers based on Publications

Braja Gopal Patra, PhD\textsuperscript{1}, Kirk Roberts, PhD\textsuperscript{2}, Hulin Wu, PhD\textsuperscript{1}

\textsuperscript{1}School of Public Health, \textsuperscript{2}School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

Introduction

Many repositories and knowledge bases have been established to ease data sharing. Most of these solutions are domain-specific and none of them recommend datasets proactively to researchers. Further, there has been an exponential increase in the number of datasets added to dataset repositories in the past last two decades. For example, in the Gene Expression Omnibus (GEO) repository, an average of 34 datasets were added to GEO daily in the last five years (i.e. 2014 to 2018). This gives a glimpse of the increasing number of datasets being made available online, considering that there are many other online data repositories as well. Naturally, it is challenging for a researcher to track repositories for potential use. The aim of this work is to proactively recommend datasets to researchers based on their profile who could potentially reuse the datasets. In the experiments, 101,279 datasets (title and summary) as of April 10, 2019, were selected from GEO. Researchers’ publications (title and abstract) were collected from PubMed after verifying collected publications with CV (obtained using a web portal) to handle the author disambiguation as shown in Figure 1.

Methods

From initial experiments, it was observed that the datasets recommended for a researcher by combining all publications together were biased towards a single research area where the researcher’s publications were most concentrated. Our insight is that the more publications a researcher has, the more interests or areas of expertise he/she likely has as well. We propose to employ the non-parametric Gibbs Sampling based Dirichlet Process Mixture Model (DPMM)\textsuperscript{1} to cluster papers into several groups of expertise. The publications and datasets were converted into vectors using TF-IDF. The publications in each publication cluster were converted into a single vector using equation: $v_c = \frac{1}{N} \sum_{p \in P_c} \lambda_{pc} v_{pc}$ and $\lambda_{pc} = \frac{1}{\exp(\frac{0.05 \times t}{1})}$, where $t$ is the difference between the current year and year of publication; and $N$ is the total number of papers in the cluster. For each publication cluster vector, top similar datasets were recommended to the researchers using cosine similarity as shown in Figure 1.

![High level architecture of researcher’s publication extraction and proposed dataset recommendation systems.](image)

Results

We asked researchers who already worked on the datasets from GEO and published papers on these to rate each retrieved dataset using one to three stars (1: not relevant, 2: partially relevant, 3: most relevant) based on their publication clusters. The system achieved partial and strict precision at 10 of 0.65 and 0.51, respectively, based on the judgments from three researchers with 53, 48 and 32 papers.

Conclusion

This work is the first step toward developing a dataset recommendation tool to connect researchers to relevant datasets they may not otherwise have been aware of. This will hopefully lead to greater biomedical data reuse and improve scientific productivity. The next goal is to identify the clusters which are not related to datasets. We plan to implement other embedding methods and test the dataset recommendation system on more users.

References


Using JATS and GraphML as a Standard Form for Clinical Practice Guidelines

Michal Patriak, Daniel R. Schlegel, PhD
SUNY Oswego, Oswego, NY, USA

1 Introduction

Clinical practice guidelines (CPGs) are distributed in a variety of formats, though rarely are they easily computer processable. This is problematic as there is a continued push for the creation of computer interpretable guidelines (CIGs), and efforts to automate this task require the input files to be in a standard, easily processable, form. Most guidelines contain textual information, formatted as a structured document with sections, headings, lists, and figures. Many guidelines also contain figures meant to illustrate portions of the diagnosis or care algorithm, often using graphs of some sort. In order to represent guidelines in a standard form, we have developed transformation guidelines for the task of translating guidelines into an XML format using the Journal Article Tag Suite (JATS) [1] and GraphML [2].

2 Methods and Discussion

The two guidelines used in the development of our transformation guidelines, ADA Standards of Medical Care in Diabetes 2017 and NCCN Non-Small Cell Lung Cancer, were distributed as PDF files. The ADA guideline takes the form of a journal supplement or book, with several chapters and using a standard layout including sections and subsections of prose along with tables and figures. Recommendations are found in highlighted boxed subsections. The NCCN guideline begins with several pages of graphs representing the diagnosis, staging, and care algorithms. The graphs are heavily footnoted, and contain links to other graphs and document sections. After the graphs is a prose discussion, including a section on recommendations. Levels of evidence are indicated in guidelines in different ways.

The document structure was encoded in the JATS format. The tag suite allows for a significant subset of HTML formatting in addition to including tags for the document structure. Tags such as ⟨publisher⟩ and ⟨title-group⟩ capture article metadata. Others such as ⟨sec⟩, ⟨list⟩, and ⟨p⟩ maintain the article hierarchy. Other tags such as ⟨title⟩, ⟨year⟩, and ⟨article-id⟩ refer to specific types of data found within the article. Tables are represented using HTML. References to figures, tables, sections, and bibliographic references use ⟨xref⟩ tags.

GraphML is used to represent graphical data, principally found in figures and NCCN graph algorithms. Raw GraphML is rather basic, allowing for the representation of nodes, edges, and simple data annotating these. The format may be extended to allow structured data inside ⟨data⟩ tags, which we have done to allow embedding JATS tags in nodes.

We have resisted re-structuring the guideline data, instead preferring to encode it in the XML format exactly as it is written or drawn. Our eventual goal is to do automatic CIG generation, and we wish for the XML documents to be as faithful to the original document as possible, even when the produced form is not as intuitive as what would be possible should this have been the original target format. The ADA guideline evidence level is shown in a lighter color after the recommendation, so we wrap it in ⟨font⟩ tags. A table is used for the NCCN evidence blocks.

This work is being used in a system for automatic clinical guideline understanding. Using the XML structure, software produces logical representations of the structural propositions in a form subsuming the Document Content Ontology, adding to it linguistic features from the underlying text.

References


---

1 This work was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R15LM013030. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

2 A DTD for the combined format is available at https://github.com/oswegonlu/JATS-GraphML
Comparing Nurse and Physician Perspectives on Health IT Adaptation

Nicole Pearl, BS1, Todd E. Tussing, DNP, RN, CENP, NEA-BC2, Esther Chipps, PhD, RN, NEA-BC2,3, Cynthia J. Sieck, PhD, MPH4, Po-Yin Yen, PhD, RN1,5

1Institute for Informatics, Department of Medicine, Division of General Medical Sciences, Washington University School of Medicine, St. Louis, MO; 2The Ohio State University Wexner Medical Center, Columbus, OH; 3College of Nursing, The Ohio State University, Columbus, OH; 4College of Medicine, The Ohio State University, Columbus, OH; 5Goldfarb School of Nursing, Barnes-Jewish College, BJC HealthCare, St. Louis, MO

Introduction

Since 2009, several policies, such as the Health Information Technology for Economic and Clinical Health (HITECH) Act, have sparked increased implementation of health information technology (IT) in healthcare settings, especially electronic health record (EHR) systems.1,2 Due to the difference in clinical duties among nurses and physicians, it is possible that they adapt, or modify existing conditions to achieve alignment,3 to new technology differently. This pilot study aims to identify factors associated with health IT adaptation, and compare nurse and physician perspectives.

Methods

We invited clinicians with varying clinical, IT, and leadership positions at a mid-west academic medical center. We asked each participant about their experiences and thoughts about health IT adaptation. All sessions were audio recorded and transcribed verbatim. Data analysis was conducted using a thematic analysis approach. Each transcript was coded by two researchers. Disagreements were discussed in meetings for consensus. Open codes were grouped into common categories from iterative discussion. Selected codes reflect the health IT adaptation processes for nurses and physicians. ATLAS.ti 8, a qualitative data analysis tool, was used for data management and analysis. Nurse and physician codes were compared to identify similarities and differences between their perspectives.

Results

Individual interviews were conducted with 13 registered nurses and 7 physicians. Preliminary results indicate themes from both groups can be grouped into 4 different categories: 1) adaptation factors, 2) outcomes resulting from the use of health IT, 3) outcomes from poor adaptation, and 4) outcomes from adequate adaptation. Both physicians and nurses feel that having off-unit training before implementation and on-site support from super users or peer experts is useful in their adaptation. Additionally, when leadership monitors health IT usage and provides feedback to users, it helps ensure the health IT is being utilized properly by physicians and nurses. Most participants would like to see increased clinician involvement in the design and implementation of health IT. Physicians specifically feel the need for more personalized training catering to individual provider roles. Physicians and nurses both recognize decreased face-to-face communication between both colleagues and patients as a result of the general nature of health IT usage, and many nurses felt frustrated regarding other clinicians’ use of copy and paste functions for notes in the EHR. Nurses and physicians also don’t have consistent access or permissions to view notes in the EHR. Ineffective and disorganized communication between providers and workarounds have resulted from poor adaption of health IT, or not using it as designed. Finding relevant patient information in the EHR has also become difficult and time consuming. Nurses and physicians expressed that when using health IT correctly, patient safety is generally increased due to decreased errors. Health IT can also lead to increased efficient communication between providers when consistently used as designed.

Conclusion

Interviews were conducted to capture health IT adaptation processes among nurses and physicians. Their perspectives were compared and grouped into 4 main themes: adaptation factors, outcomes of the general nature of health IT, outcomes from poor adaptation, and outcomes from adequate adaptation. Future studies should further investigate how nurses and physicians utilize health IT differently to allow clinical and IT leaders to facilitate innovative plans to increase adaptation across all provider roles as efficiently as possible.

References

Adapting a Patient-Reported Outcome App and Practice Model into Primary Care for Treatment of Asthma

Sofia Perez,1 Anuj K Dalal, MD,1,2 Jessica Sousa, MSW, MPH3, Robert S. Rudin, PhD3

1Brigham and Women’s Hospital, Boston, MA 2Harvard Medical School, Boston, MA 3RAND Corporation

Introduction: Asthma is largely controllable through symptom monitoring and treatment adjustment to avoid clinical deterioration and emergency care. Useful tools for symptom monitoring, such as patient reported outcome (PRO) questionnaires, have the potential to improve adherence to guidelines, but have not been widely adopted. The purpose of this study is to adapt a mobile health (mHealth) app and practice model that leverages electronic PRO questionnaires for asthma symptom monitoring that was previously developed for subspecialty care, for use in a primary care setting.

Methods: We are working with 4 ambulatory clinics affiliated Brigham and Women’s Hospital in Boston, MA to adapt our intervention to primary care. To guide our design process, we are using the NASSS framework, developed specifically to ensure scalability and sustainability of the intervention1. We are engaging multiple stakeholders during every step of the adaptation process: primary care providers (PCP), patients, nurses, clinic leadership, venders, and administrative staff who are responsible for technical infrastructure. We are employing a user-centered design approach (design sessions, low-fidelity/high-fidelity wireframes, surveys) to identify user requirements; understand workflow for patients, nurses, and clinicians; and determine how to integrate our intervention with our vendor electronic health record (Epic Systems, Inc.). Design sessions comprised of interview questions which were recorded, transcribed and summarize in consensus of two members of the research team.

Results: Based on 7 individual design sessions (6 physicians, 1 patient) conducted to date, we have identified several new requirements (Table 1) for scaling our intervention to primary care.

Table 1: User Requirements by Intervention

<table>
<thead>
<tr>
<th>Intervention Component</th>
<th>Previously Developed</th>
<th>New Requirements for Primary Care</th>
</tr>
</thead>
</table>
| Invitation             | • Patients receive a letter from their PCP inviting them to participate  
                          • Research staff calls patients and schedules in-person meeting. | • To facilitate enrollment and improve scalability, patients should be able to send a text message or email directly to address or phone number directing them to REDCap enrollment link |
| Weekly questionnaire and option to request a call-back from a nurse | • Patients receive weekly 5-item questionnaire and can request a call-back if symptoms deteriorate.  
                          • Nurses receive an email notification to login to dashboard | • Patients without active symptoms should respond to just 1 question  
                          • Nurse receives message in EHR ‘inbasket’ when patients request a call-back and should follow an institutionally supported triage protocol |
| Patient self-management | • Patients can view their previous reported symptoms on a graph | • Patients can view/enter notes, peak flows, triggers. |
| In-person visit        | • PRO data is available in a recent physician note prior to a visit | • Clinicians should be able to access PRO quickly from the EHR (e.g., one-click) and receive an alert when PRO data is available to review  
                          • Patients should receive a reminder about in-person visits through the app and are encouraged to review their symptoms with physicians |

Conclusion: Design sessions require feedback from all stakeholders as they play a crucial role in determining new requirements for integration. Adaptation of a relatively simple mHealth app and practice model for asthma symptom monitoring in the primary care setting still requires an extensive design process due to the need for integrating with clinical processes, leveraging available health system infrastructure, and reconciling many user requirements. The NASSS framework has been a helpful tool that has enabled us to prioritize requirements for primary care.

References
INTRODUCTION

REDCap uses user-defined fields and instruments to create custom databases. Records can be imported to a REDCap database from a variety of sources, including a comma-delimited (CSV) file or an excel spreadsheet. Current methods for preparing records for upload to REDCap generally involve manual transformation of data. At Memorial Sloan Kettering, we are working on efforts to remove manual steps when entering data into REDCap projects. REDCap Linter is a wizard that walks a user a multistep process of rapidly cleaning up errors in source data before loading it into REDCap. First, the user matches spreadsheet headers with REDCap fields and cleans up errors found in any column’s data. Many tasks like aligning REDCap fields with spreadsheet column headers and matching REDCap fields with permissible values and their corresponding column in a datafile are tedious and error-prone. Some fields such as checkboxes and dropdowns require a set of permissible values. For example, a study may have a ‘Demographics’ instrument with a field for ‘Gender’ that contains the set of permissible values (Male | 1, Female | 2). These fields need to be recoded with appropriate permissible values before upload. REDCap Linter 1.0 generates a cleaned csv file that is ready for upload to REDCap. However, there were still some additional features that were lacking if this cleaning (aka linting) process were to be less manual.

NEW FEATURES

With REDCap Linter 2.0 we build on our previous submission by introducing new features that enable the continuous flow of data into REDCap projects. REDCap Linter 2.0 introduces:

1. The ability to export and import a mappings file that stores information connecting spreadsheet columns to REDCap fields. The mappings file also stores steps used to clean up errors found in column data.
2. An API to call on REDCap Linter services programmatically.

The ability to save project mappings is an important feature for projects that receive updated spreadsheets in a consistent format. The feature allows a quick load of the new spreadsheet without having to manually re-map the columns again to the fields in REDCap. If additional errors not already stored in the mappings file are found in the new spreadsheet data, they can be added to mappings file.

The API to call Linter allows automation. Input parameters would include mappings file, data dictionary and spreadsheet data. Automation pipelines can create the mappings file for each spreadsheet sent to the Linter and receive back a cleaned CSV file for upload into REDCap.

CONCLUSION

At Memorial-Sloan Kettering, we are working on solutions to dynamically pull data into REDCap projects, however REDCap Linter is a general approach and is agnostic to the data source. REDCap Linter is open-source and we welcome any contributions from the community.

REFERENCES


Kevin Peterson, MS1,3, Hongfang Liu, PhD2

1 Division of Information Management and Analytics, Mayo Clinic, Rochester, MN
2 Department of Health Sciences Research, Mayo Clinic, Rochester, MN
3 Bioinformatics and Computational Biology Program, University of Minnesota, Minneapolis, MN

Introduction. Despite ongoing standardization efforts in healthcare, the goal of consistent data semantics has remained elusive. First, by their nature standards are often focused on elements of the domain that are broadly agreed-upon, and thus often struggle to account for the idiosyncrasies of specialized use cases, forcing users into expensive and time-consuming workarounds. Second, while advances in data-driven techniques have shown promise for eliciting robust semantic models from existing Electronic Health Record (EHR) data, blueprints on how to align the “top-down” development approach employed by standards organizations with “bottom-up” data-driven approaches are sparsely described in literature. Finally, downstream applications often have very pragmatic data needs. Standards that are too complex or abstract lead to implementation challenges and high costs – or the abandonment of standards altogether in favor of more one-off, pragmatic approaches. Semantic models must be easily consumable and able to be integrated into existing healthcare systems. In this work we propose three components of a framework to combine data-driven techniques with domain standards to produce interoperable semantic models. The components of the framework are enumerated below. Our intent for this work is to outline the high-level features of our system, setting the scope and direction for future research efforts in this area.

Component 1: Extract and standardize clinical concepts from narrative text.
Description: Using data-driven Natural Language Processing (NLP) techniques including sublanguage analysis and the transformation of text into SNOMED CT Expressions, the goal of this component is to extract and catalog concepts from clinical notes. This will output a data-driven “glossary” of how clinical concepts are defined in practice.

Component 2: Mine domain models for clinical concepts by combining data-driven and standards-based approaches.
Description: By leveraging ontology extraction techniques with Web Ontology Language (OWL) models, this component will mine domain models specific to the concepts extracted in Component 1. Standard, expert-curated domain ontologies will be used to bootstrap our data-driven approach, as well as to contrast the data-derived results with standard representations to quantify the differences – bridging the gap between standards and data.

Component 3: Automate the creation of Fast Healthcare Interoperability Resources (FHIR) Profiles based on clinical concepts (Component 1) and associated ontologies (Component 2).
Description: FHIR Profiles are an effective way of creating enforceable, unambiguous data contracts between applications. This component uses the inputs of Components 1 & 2 and outputs semantic models as FHIR Profiles.

References
Agent Based Model to Evaluate Simple Behavioral Changes on Individual Health Over Time

Andrew B. Phillips, PhD, RN\textsuperscript{1}, Margarita Sordo, MSc, PhD\textsuperscript{1,2,3}
\textsuperscript{1}MGH Institute of Health Professions, School of Nursing, Boston, MA, \textsuperscript{2}Department of Medicine, Brigham and Women’s Hospital, \textsuperscript{3}Harvard Medical School, Boston, MA

Introduction
For any disease, symptom, or condition, the underlying goals are to identify it as early as possible, understand its pathology, and employ an intervention that ameliorates symptoms, and/or eliminates the disease/condition (cures it). Health interventions are often born by the patient – changes in diet, exercise, smoking, routine monitoring, etc. The ability to improve our understanding of patient behaviors can ultimately lead to improved interventions with increased rates of success. Agent Based Models (ABM) represent an approach to modeling complex patient behaviors over time. An ABM is defined by individual agents (typically individuals) with their own attributes (descriptors) and behaviors (actions) residing within a simulated environment. The simulation produces an emergent behavior of the system based on the interaction of the defined attributes and interactions of its individual agents and the environment. ABMs can see the cumulative effects of negative or positive individual (patient) behaviors over time.

Methods
We selected exercise as an example as it represents a significant contributor to preventive health. The model used a simple set of rules. Our population mirrored the US population by age and sex. Each time point in the model represented a week and the simulation stopped after 5000 weeks (approximately 96 years). The ABM was guided by the following assumptions:

A – Exercise positively impacts health
B – Health impacts desire to exercise
C – Desire to exercise impacts exercise
D – Reduced impact of age over time on desire to exercise and overall health and exercise intensity, including probability of death.
E – Random element of overall health. For example, an individual could develop a chronic disease which impacts desire to exercise.

General Model Assumptions
1 - Individual health decays over time
2 - Individual Life expectancy per CDC
3 - Desire to exercise declines with age

Results

| Vigorous exercise increases health score, even over the extended period of time in the model. Other levels of exercise still have positive affect compared to no exercise, but health score decreases over model time frame. | Even though desire to exercise declines with age, the drop over time is influenced by health level. Good health will slow down the decline, whereas poor health will accelerate it. |

Discussion
Individual behavior is an important determinant of health. Changes, even small ones, to our individual behavior, in the form of simple actions, i.e. physical activity, can have positive altering effects over time. ABM models provide an opportunity to apply new tools to simulate and model individual behavior that evaluate the consequences of simple behaviors over time. Our model highlights two important aspects of individual behavior: a) simple actions have a direct cumulative effect over time; with b) cascading, feedback loops as part of a chain of cause-effect. As shown by our model, simple actions chained into cascading feedback loops, promote other actions/behaviors with their own consequences, while allowing agents to reach their own equilibrium. Under this approach, we can assess the short- and long-term impact of behavioral health interventions, and their cascading effects. Physical activity is a simple, modifiable behavior with a positive impact on health. Good health has a potential positive impact on desire to exercise, which in turn has an impact on the actual intensity of exercise. Although, conceptually simplified, our model captures the essential nuances and effects of a simple behavioral intervention in the overall well-being of an individual, that, when sustained, can have lasting benefits over time.

References
Clinical Decision Support Systems in Pediatric Surgery: a Scoping Review

Michael R. Phillips, MD1, Christopher J. Tignanelli, MD2, Saif Khairat, PhD, MPH1, Arlene E. Chung, MD, MHA, MMC1
1University of North Carolina, Chapel Hill, NC, USA; 2University of Minnesota, Minneapolis, MN, USA

Introduction
Electronic clinical decision support systems (CDSS) are systems designed to facilitate decision-making during the diagnosis and treatment of a medical condition. Pediatric surgeons treat a variety of surgical conditions in neonates to young adult patients. During episodes of care, pediatric surgery patients are also exposed to many different providers in outpatient and inpatient settings. This heterogeneity of patient demographics, providers, and clinical settings makes effective CDSS critical for high quality care. However, little is known about the usability or effectiveness of CDSS in pediatric surgery.

Methods
To review the literature in CDSS for pediatric surgery, the Pubmed database was searched using a combination of MESH terms related to pediatrics, children, surgery, and clinical decision support or clinical decision support systems. Studies were included if they met the following criteria: published in English, journal articles or conference papers that described the prospective use of any electronic CDSS in children (≤18 years old) with surgical conditions. Exclusion criteria were articles without electronic decision support, and articles on clinical diagnostic scoring systems. Two reviewers performed initial abstract screening, and all authors reviewed included articles for clinical relevance and findings.

Results
There were 69 articles that met the search criteria. After initial screening, 12 papers were included (inter-rater reliability between two reviewers=93.9%). Upon final review of the 12 papers, four additional papers were unanimously excluded, two for the retrospective use of the CDSS tool, one for duplicate data from another included study, and one that focused on adults treated in a pediatric intensive care unit. Eight papers were included in the final scoping analysis. Papers analyzed the following topics: order sets for appendicitis (2), medication selection and dosing assistance (2), electronic triage tool for abdominal pain (1), computer analysis of orthopedic conditions compared to expert opinion (1), alerts to assist with anesthetic management of traumatic brain injuries (1), and a home monitoring system for pediatric liver transplant patients (1). None of the papers examined the effect of the CDSS system on patient outcomes. Five of the papers reported user satisfaction or acceptance with rates of 69-100% and 20-30%, respectively. Of the studies examining satisfaction and acceptance, one examined electronic antimicrobial stewardship, one provided diagnostic and therapeutic recommendations during anesthesia for neurosurgical procedures, one compared an ad hoc order set to a systematically developed order set, one embedded clinical advice within an order set, and one evaluated acceptance of medication dosing alerts within a computerized order entry system. All of the studies that examined user satisfaction reported increased satisfaction with clinical decision support. Of the two studies examining practice changes, only one paper could connect practice patterns directly to CDSS. This study demonstrated similar changes in practice between pediatric and non-pediatric providers, but that pediatric trained providers were more likely to ignore an alert and less likely to cancel an order for medications outside the recommended dosing range. Three studies reported the diagnostic accuracy of CDSS (37.5%), including: one study comparing electronic vs. expert evaluation of clubfeet deformities (86% concordance), one examining the rate of detection of brain-injured patients to receive anesthetic treatment alerts during surgery (79%), and one measuring the rate of correct triage for patients with abdominal pain in the emergency room using a mobile app (70%).

Conclusion
The effectiveness of CDSS tools in pediatric surgery is an emerging area of research. This study is limited by the heterogeneity of the CDSS tools studied and the inconsistent reporting of variables examined. There was no study that examined the effect of the CDSS tools on patient outcomes. Studies examining user satisfaction demonstrated improved scores compared to historical practices. It appears that improving CDSS tools in pediatric surgery may increase user satisfaction with the electronic health systems; however, further research is needed to determine which types of CDSS provide the most improved satisfaction and make the most impactful changes to clinical practice.
Implementing Data Governance for Precision Health

David J. Pierce, BA¹, Kun Huang, PhD², Paul Biondich, MD MS³, Shaun Grannis, MD MS⁴, Burke Mamlin, MD⁵, Peter Embi, MD⁶

¹Indiana University, Indianapolis, IN; ²Indiana University, Indianapolis, IN; ³ Indiana University, Indianapolis, IN; ⁴ Indiana University, Indianapolis, IN; ⁵ Indiana University, Indianapolis, IN; ⁶ Indiana University, Indianapolis, IN

Introduction

The Indiana University (IU) Precision Health Initiative is a five year multidisciplinary project that aims to: cure at least one cancer, cure one childhood disease, prevent one chronic disease, and prevent one neurodegenerative disease. In order to achieve these goals, data from multiple systems and sources are being combined to support various research projects.

The research projects supported by the initiative require clinical, genomic, environmental, and other data such as information about biospecimens. The required data is spread across multiple systems and is owned by various data providers. In order to support the research project, an effective and transparent process had to be implemented that would allow for this type of broad data sharing.

Methods

Initially, the initiative approached data sharing by working through each data provider’s process. This approach resulted in delays in accessing the data and was too onerous for all parties involved. Instead of working through each data provider’s process, the initiative set forth a set of guiding principles for data governance. Four committees (Figure 1) were chartered with specific mandates to implement and monitor progress against the guiding principles.

![Data Governance Committee Structure](image)

The Executive Committee sets overall data policy for the initiative and resolves high level policy issues. The Data Provisioning Committee ensures that all required agreements are in place to allow data sharing. The Security and Compliance Committee defines security standards that meet all data providers’ and regulation requirements for data security. The Data Quality Committee is responsible for setting data standards and driving data quality improvement projects.

Each committee has representation from each data provider, which increases engagement from the data providers. The chairs of the sub-committees sit on the Executive Committee and provide regular reports on progress and challenges.

New projects are evaluated by the sub-committees and when necessary are escalated to the Executive Committee. Project submission templates were created for the researchers that guide them through the submission process to reduce delays.

Conclusion

Strong data governance is required in order to support data sharing across multiple institutions at scale. Standardizing the approach to data governance across the data providing organizations has allowed the initiative to more efficiently respond to new project requests thus increasing the speed of research.
Augmenting Quality Assurance Measures in Radiation Oncology with Machine Learning

Malvika Pillai1, Karthik Adapa, MPH, MBBS1

1Carolina Health Informatics Program, University of North Carolina, Chapel Hill, NC

Introduction

According to the current nationally regulated approach to quality assurance (QA), the physicist or dosimetrist manually chart checks multiple metrics for each patient before administering radiation therapy. Chart checking is time intensive, involves high cognitive burden, and is only 60% effective in detecting high severity incidents (1). Presently, in-house QA tools augment chart checking by automating standardized second checks, thus improving efficiency and reducing cognitive workload (2). The research objective is to improve the effectiveness of chart checking and reduce cognitive workload by flagging cases requiring additional manual and cognitive scrutiny. The analysis demonstrates a classification algorithm that categorizes the radiation treatment plans as hard, medium, and easy based on cognitive scrutiny involved.

Methods

We collected data from pre-treatment and weekly chart checks for treatment plans of each patient, encompassing all cancer sites. Data used in this analysis was collected from August 2018 to June 2019. A clinician and a physicist selected data attributes based on clinical relevance, contribution to plan complexity, and quality assurance metrics. The degree of difficulty of treatment plans, an outcome variable, was collected as a subjective rating by physicists on a scale of 1-10 from physics pre-treatment chart checks. 898 patient plans were used as a training set, where a patient could have multiple plans. All analysis was done using Python sklearn packages. A majority voting classifier was used in predicting the degree of difficulty of radiation treatment plans and classifying them as easy (1-3), medium (4-6), and hard (7-10). The class distribution of the dataset was: 52% easy, 38% medium, and 10% hard. The majority voting classifier consisted of 5 algorithms: support vector machine (SVM), decision tree, random forest classifier, adaboost, and k-nearest neighbor (KNN) where k is 2. Before predicting the degree of difficulty of plans, forward feature elimination was utilized to incorporate only variables of importance into the predictive algorithms (3). 17 attributes were used as input, and forward elimination was used for decision tree, random forest, and adaboost. The common variables selected for decision tree, random forest, and adaboost were: age, number of fractions, number of beam sets, number of images, and plan check completion time (ie 12:34 pm). Random forest and adaboost additionally selected organ count and site name as variables, while adaboost also used dose per monitor unit (MU).

Results

As the project is in progress, a blind test set has not been used yet to validate results. The training set of 898 patient plans was split into 80% for training and 20% for testing. Accuracy with 95% confidence intervals (CI), positive predictive value (PPV), and sensitivity were collected for each of the algorithms in the majority voting classifier (Table 1).

Table 1. Results by Algorithm.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>SVM</th>
<th>Decision Tree</th>
<th>Random Forest</th>
<th>Adaboost</th>
<th>KNN</th>
<th>Majority Voting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.97</td>
<td>0.89</td>
<td>0.99</td>
<td>0.64</td>
<td>0.87</td>
<td>0.95</td>
</tr>
<tr>
<td>95% CI</td>
<td>[0.95, 0.99]</td>
<td>[0.84, 0.94]</td>
<td>[0.98, 1.00]</td>
<td>[0.58, 0.71]</td>
<td>[0.82, 0.92]</td>
<td>[0.92, 0.98]</td>
</tr>
<tr>
<td>PPV</td>
<td>0.98</td>
<td>0.92</td>
<td>0.99</td>
<td>0.65</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.96</td>
<td>0.84</td>
<td>0.99</td>
<td>0.56</td>
<td>0.76</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Conclusion

We present a majority voting classifier in predicting difficulty of radiation treatment plans and flagging cases requiring extra cognitive scrutiny. The project utilizes clinical, plan complexity and QA data in generating knowledge about components of the radiation therapy plan that contribute to plan difficulty and in predicting the effect of plan components on cognitive workload. Further research on the project will consist of addressing limitations such as missing data, class imbalance, and using another classifier instead of adaboost.

Acknowledgement: First author funded by National Library of Medicine through grant number 1-T15-LM012500.

References


1709
Developing an Electronic Chart Review Tool to Identify and Assess Diagnostic Errors in the Acute Care Setting

Nicholas Piniella¹, Kerrin Bersani¹, Kumiko Schnock, PhD, RN¹, Pamela Garabedian, MS¹, David W Bates, MD, MSc¹,², Jeffrey L Schnipper, MD, MPH¹,², Anuj K Dalal, MD¹,²
¹Brigham and Women’s Hospital, Boston, MA; ²Harvard Medical School, Boston, MA

Introduction: Diagnostic error in acute care represents an unresolved safety issue: error rates range from 4.8 to 49.8%. Though chart review tools to determine the presence of diagnostic error have been validated in the ambulatory setting, these have not been validated for acute care. The purpose of this study is to describe our approach to developing a chart review tool that can be used to assess the presence of diagnostic errors in the acute care setting, as well as their potential for harm and categorize types of failures in the diagnostic process.

Methods: We used REDCap, an electronic data capture tool, to design and develop a chart review tool that clinician adjudicators could use to determine the presence of diagnostic error and diagnostic process errors in acute care. After conducting a literature review, our team decided to utilize both the Safer Dx Instrument and the Diagnostic Error Evaluation and Research (DEER) Taxonomy. Next, we gathered feedback from clinicians and researchers with expertise in studying diagnostic errors to adapt these tools for use in the acute care setting. Specifically, we made slight wording changes to the Safer Dx Instrument questions to assess the diagnostic process from admission to discharge. We adapted the DEER Taxonomy to capture specific types of failures in the diagnostic process during hospitalization: most notably, we added 2 domains, one to capture failures in care team communication and collaboration, and the other to identify issues related to patient experience. Finally, we asked physicians to use our tool to conduct electronic chart review of selected cases of potential diagnostic error. Based on their feedback, we iteratively refined our tool and chart review process, addressing issues that were identified.

Results: The final chart review tool (Table 1) has 4 sections that ask clinician adjudicators to identify and assess diagnostic errors during acute care. To create a workflow that would not be prohibitively long for reviewers, we used branching logic, based on the admission and discharge diagnoses and whether appropriate treatment was initiated within 24 hours, to guide the reviewer through the chart review process. For example, if there was preliminary evidence of a delayed diagnosis (i.e., discharge diagnosis substantially different than the admission diagnosis), then the tool encouraged the reviewer to focus on timestamps for diagnostic tests, labs, medications, and consultation orders. We also decided that reviewers would identify and enumerate all diagnostic errors that occur during a patient’s hospitalization, and identify DEER failure points in all reviewed cases, regardless of the presence of diagnostic error.

Conclusion: We created an electronic chart review tool and process to facilitate identification of cases of diagnostic error in acute care. Our next steps are to validate this chart review tool by: 1) retrieving cases of potential diagnostic error from the EHR using a variety of triggers (e.g., transfer to intensive care); 2) having two clinicians independently review charts of previously hospitalized patients who have one or more trigger to determine the presence of diagnostic error; and 3) calculating the sensitivity, specificity, and positive predictive value for trigger positive cases.

Table 1. Sections of the Chart Review Tool for Identifying and Assessing Diagnostic Error in Acute Care

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safer Dx Instrument</td>
<td>- Reviews answer 11 questions on a 6-point scale to assess the appropriateness of the diagnostic process</td>
</tr>
<tr>
<td></td>
<td>- The tool calculates a diagnostic error score, indicating the risk of a diagnostic error in the episode of care</td>
</tr>
<tr>
<td></td>
<td>- Reviewers use the error score and their clinical expertise to determine if an error was present</td>
</tr>
<tr>
<td>Number of Diagnostic Errors</td>
<td>- Reviewers enter the number of diagnostic errors, and a description of each error</td>
</tr>
<tr>
<td>Adverse Events</td>
<td>- Reviewers determine if there was actual harm or the potential for harm attributed to the diagnostic error</td>
</tr>
<tr>
<td></td>
<td>- Reviewers rate the severity, preventability, and ameliorability of the adverse event</td>
</tr>
<tr>
<td>DEER Taxonomy</td>
<td>- Reviewers determine if any of 41 possible failure points were present across 8 domains: Access/Presentation, History, Physical Exam/Assessment, Diagnostic Test Ordering, Performance, and Interpretation, Diagnostic Information and Patient Follow-up, Subspecialty Consultation/Referral, Healthcare Team Communication &amp; Collaboration, Patient Experience</td>
</tr>
</tbody>
</table>

Acknowledgements: The PSLL project is supported by funding from AHRQ (R18-HS026613).

References:

Improving Post-traumatic Stress Disorder (PTSD) Detection and Treatment After Traumatic Brain Injury (TBI)

Christie Pizzimenti, PhD1, Claire Ramey1, David Dorr, MD, MS1, Kari A. Stephens, PhD2
1Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR; 2Department of Psychiatry & Behavioral Sciences, University of Washington, Seattle, WA

Introduction

An estimated 2.8 million Americans sustain a traumatic brain injury (TBI) each year1 with an elevated (5.8 times) relative risk for Post-Traumatic Stress Disorder (PTSD)2. Estimates of epidemiological rates of PTSD following TBI range between 7 and 20%3, but healthcare organizations struggle to identify and treat PTSD. Subsequently, the costs of repeat hospitalizations and emergency department visits for PTSD-related symptoms go unmitigated. Electronic health record (EHR) data can be used to help identify patients with a high risk of developing PTSD following TBI and promote follow-up for appropriate treatment. We therefore explored the possibility of using EHR system data to understand the frequency of PTSD after TBI in treated populations at Oregon Health & Science University (OHSU).

Methods

ICD-10 codes were used to first identify patients diagnosed with TBI within the EHR, then patients also diagnosed with PTSD. Patients without a PTSD diagnosis, but with symptoms extracted from the EHR were also identified via standard PTSD rating scales (i.e., PTSD Check List –C and -5), given their high concordance with a PTSD diagnosis. TBI patients with subsequent encounters were identified via hospital admissions and appointments after the date of injury to assess opportunities to provide follow-up care.

Results

At OHSU, 3,475 patients with TBI were seen over the last 3 years. PTSD diagnoses were identified via ICD-10 codes in 49 patients (1.4%) and PTSD diagnoses were likely in 195 patients (5.6%) via positive screens. Indications of PTSD via screeners and not ICD-10 codes most closely represented expected epidemiological approximation of the rate of PTSD subsequent to TBI. It was also found that TBI patients were likely to seek care after TBI, with 45% returning via hospital admissions and 67% via office visits after the injury date.

Discussion

PTSD diagnosis and screening rates under-detected rates of PTSD in a TBI population, likely indicating better screening and treatment for PTSD is needed for patients with TBI. However, despite the low detection rate, EHR data, particularly via screener values, may be used to identify patients with TBI who are likely to be experiencing PTSD and likely to return to the hospital after TBI. PTSD screener values provide valuable information needed to facilitate effective treatment for TBI and PTSD. Additional refinement of this approach should target investigating common comorbid conditions and use of predictive analytics to drive better care.

References

Harmonizing Oncology Data from Multiple Institutions with the Observational Medical Outcomes Partnership’s Common Data Model

Joseph M Plasek, PhD1, John A. Weissert1, Kourosh Ravvaz, MD, PhD, MPH1
1Aurora Research Institute, Advocate Aurora Health, Milwaukee, WI, USA

Introduction
Semantic harmonization of disparate electronic health record (EHR) expedites multi-institutional clinical research as all data elements have a shared meaning. Common data models (CDM) with controlled vocabularies and a standardized structure provide a framework for achieving semantic harmonization. The Observational Medical Outcomes Partnership (OMOP) CDM enables systematic analysis of disparate data sources by transforming data into a common format with a common representation1. Limited prior use of OMOP CDM in oncology included cohort identification of prostate cancer2, among others. Using OMOP CDM, we aimed to test the feasibility of harmonizing multi-institutional non-muscle invasive bladder cancer (NMIBC) data, a disease noteworthy for a patient population with heterogeneous risk profiles and clinical outcomes. In doing so, we identified opportunities to improve the OMOP CDM for analyzing bladder cancer EHR data. Bladder cancer is the ninth most common cancer of which 70% of new cases present as NMIBC3.

Methods
From 1991 - 2017, two retrospective datasets containing 509 and 1331 patients with NMIBC were independently curated from the University of Wisconsin Madison and Aurora Health Care, respectively. We utilized OMOP CDM version 5.3.1 to transform the heterogeneous longitudinal datasets (consisting of cancer registry, pathology, and billing data) into a standardized format to enforce consistent structure and semantics.

Results
The OMOP CDM was insufficient for representing some aspects of the NMIBC data. Gaps identified revolved around the generalizability of OMOP CDM to represent oncology specific attributes, complex drug regimens, and time to outcome. Gaps: 1) The structure and/or vocabulary for storing oncology specific attributes (e.g., number of tumors, tumor size). These tumor characteristics were dispersed across multiple tables (e.g., condition occurrence, measurement, observation) without enough detail and linkages. 2) OMOP CDM lacked structure to express immunotherapy regimen elements (e.g., induction vs maintenance, cycle number) for drugs such as Bacillus Calmette-Guerin. 3) OMOP CDM does not allow for an explicit representation of time between recurrences for an individual patient. To address these gaps, we adapted existing concepts for tumor specification, repurposed partially related fields for drug regimens, and created temporary categorical representations of otherwise continuous data.

Conclusion
Despite the increasing attention and rigor of OMOP CDM, the growing multiplicity of EHR data sources, structures and thereby semantic harmonization may remain challenging in the foreseeable future. To improve OMOP CDM, we advocate for the development of more appropriate oncology vocabularies and table structures that can adequately represent continuous and temporal oncology data.

References
Using “Big Data” to Determine Prevalence of Side Effects of Chemotherapy in Breast Cancer Patients

Lu Pu, MD, PhD¹, David C. Kaelber, MD, PhD, MPH²
¹School of Medicine; ²Depts of Internal Medicine, Population and Quantitative Health Sciences, The MetroHealth System, Case Western Reserve University, Cleveland, OH

Introduction

One of the key factors affecting the care and mortality of breast cancer patients is the cardiotoxicity of chemotherapies. The cardiotoxicity of antimetabolites including 5-fluorouracil and capecitabine has been demonstrated in case reports, randomized control trials, and prospective and retrospective studies (1, 2), with incidence rates ranging from 1% to 68% (3). Aggregated and de-identified population data has been used for post-market surveillance of other non-chemotherapy medications (4). Here we present the first proof-of-principle study on post-market surveillance of cardiotoxicity of breast cancer therapies using the OHDSI (Observational Health Data Sciences and Informatics Collaborative)/OMOP (Observational Medical Outcomes Partnership) common data model (CDM).

Methods

De-identified health records of 2,740,000 breast cancer patients were obtained from five OMOP CDM-based electronic health record databases that were part of OHDSI. Data analysis occurred over a course of 10 hours using Atlas tools on top of the OHDSI databases (atlas.ohdsi.com). Diagnoses (breast cancer and medication side effects) were searched using, SNOMED terms and medications where searched using RxNorm for drugs.

Results

The incidence rate of left ventricular dysfunction (LVD) in breast cancer patients treated with antimetabolites was 2-3 times higher than left ventricular dysfunction with other therapies (Figure 1).

Discussion/Conclusion

Here we demonstrate that a medical school student (author LP) with limited exposure and training to large EHR data sets quickly (~10 hours of work) used tools to easily explore large data trends, in this case side effects related to chemotoxicity among breast cancer patients. Especially for relatively rare conditions this approach is attractive because other methods (case studies, randomized control trials, and traditional cohort studies) are not effective or efficient at assessing the overall frequency of correlation of rare events.

Large EHR databases, such as OHDSI and related analytics tools can be used to efficiently and effectively assess correlation between various chemotherapies and their potential side effects.

References


Figure 1. Incidence rates of LVD in patients with breast cancer and treated with: taxol (Taxanes, eg., paclitaxel), tam (endocrine therapy, eg. tamoxifen, anastrozole, letrozole), radiotherapy, her2 (HER-2 targeting therapies, eg., trastuzumab, pertuzumab), dox (anthracyclines, eg., doxorubicin, epirubicin), cisplatin (alkylating agents eg., cisplatin, cyclophosphamide), cdk (cyclin-dependent kinase 4/6 inhibitors, eg., palbociclib, ribociclib), and antimetab (antimetabolites, eg. 5-fluorouracil, capecitabine). Significance (by Chi-square test) is indicated (*: *p<0.05, **: *p<0.01).
A Modular Web Application Offers Flexibility and Adaptability to Cope with Frequent Structural Updates in Pathology Cancer Reporting

Zhenhong Qu, MD, PhD 1; Jason Guo Jin 2, Keran Zhao 3, Elaine Qu 4, Kimberly Pieters 1.

1. Department of Pathology, Beaumont Hospital, Royal Oak, MI.
2. University of Michigan, Ann Arbor, MI.
3. Department of Information and Decision Sciences, University of Illinois, Chicago, IL.

Introduction/Background
The American College of Surgeons Commission on Cancer (ACoS-CoC) mandates that pathology reports of cancer must include essential data elements and follow specific formats. An online application with dynamic forms has been an effective tool for pathologists to file cancer reports in compliance with the ACoS-CoC’s mandates. However, frequent updates of the reporting standards require corresponding back-end structural changes of the Web forms as reporting templates, which for more than a decade remains a daunting challenge in keeping the reporting templates up-to-date. No effective solution has been identified. Here we introduce a new method by which the users can 1) update/create reporting templates by combining dynamic form units as software modules on the Web via a simple drag-and-drop operation, and 2) file tumor reports with the updated reporting templates. This modular method effectively overcomes the challenge by greatly improving the flexibility in building cancer reporting templates and keeping them up-to-date when new reporting standards and requirements arise.

Method and Results
Our survey of more than a hundred of different tumor synoptic report forms by four different international professional organizations has identified a small number of frequently employed functionalities/features of Web forms (e.g. drop-down list, checkbox, free text box, etc.) that can meet nearly all format needs form reporting forms. Accordingly, we created Web modules for each of these functional form units as JavaScript Object Notation (JSON) Schema using ReactJS, a popular JavaScript library maintained by Facebook, Inc. (Menlo Park, California 94025). The JSON Schema defines Web forms. The final application uses JSON and HTML to create dynamic forms on the Web. The users can selectively upload multiple individual JSON files (i.e. form modules) via drag-and-drop on the Web page to create customized Web forms on the fly. The combined JSON file for a Web form can then be downloaded for future use as a reporting template or again as a module to merge with other form modules for updating or making new reporting templates. The resultant Web forms (templates) are used for online tumor reporting by pathologists. This platform is accessible at: http://www.essentialpathology.info/index.html?home by clicking icon “Create template …”

Conclusion and Discussion
This new modular method allows users to create and update tumor reporting forms easily in a Web browser without demanding programming skills of most users. It effectively overcomes the obstacle to timely updates of tumor-reporting standards mandated by ACoS-CoC. It is a simple method which solves a big chronic problem in the standardization of cancer reporting in pathology.

References
DSL-TEACH: Data Science Literacy Training to Enhance Approaches for Clinical decision-making in Healthcare

Samir Rachid Zaim, BA1,2, Ahyoung Amy Kim, BS3, Colleen Kenost, EdD1,2, Helen Hao Zhang, PhD3,6, Yves A. Lussier MD,1,3,7,8 Vignesh Subbian, PhD4,5

1Center for Biomedical Informatics & Biostatistics, 2Dept. of Medicine, 3Graduate Interdisciplinary Program in Statistics, 4Dept. of System and Industrial Engineering, 5Dept. of Biomedical Engineering, 6Dept. of Mathematics, College of Sciences, 7BIO5 Institute, 8UA Cancer Center,
The University of Arizona, Tucson, Arizona

Problem Statement

Prior studies1-3 surveying physicians indicated that many commonly-used statistical metrics, such as positive predictive value (PPV) or relative risk (RR), are not well understood, despite their coverage during medical training and subsequent use in clinical care. Indeed, Clinical Decision Support Systems (CDSS) implemented in Electronic Health Records (EHRs) are complex, dynamic data-driven healthcare tools that implicitly require a basic level of statistical literacy (e.g., PPV) for proficient development and utilization. However, new tools stemming from data science (i.e., case-based reasoning, machine learning-anchored practice-based evidence) are advancing the field of clinical decision-making4-5, demanding greater awareness for their utility and application in clinical care. Further, practice-based evidence derived from machine learning applied to big clinical data poses new challenges for clinicians. For example, the seminal NEJM on calculating risks of therapy from EHRs presented four relative risks without documenting their interdependencies, leading to a challenging interpretation6.

DSL-TEACH: Proposed Framework and Learning Modules for Promoting Data Science Literacy

To improve awareness and clinical utility of forthcoming data science tools in healthcare, we propose DSL-TEACH (Data Science Literacy Training to Enhance Approaches for Clinical decision-making in Healthcare), a framework for promoting literacy in data science and predictive analytics applied to clinical decision making. For example, components will include i) understanding data pre-processing and biases, ii) confounder-controlled observational study design, iii) strengths and limitations of feature curation and selection, iv) strengths and limitations of machine learning models, as well as v) fundamentals of probabilistic and case-based reasoning (e.g., understanding of risk reporting). Our current prototype is based on the above-mentioned components; however, we are in the process of conducting a needs assessment to better cater our framework to meet the specific needs of our audience. Once finalized, the polished prototype will be deployed as online learning modules, hosted via REDCap, which will be freely available at CB2.uahs.arizona.edu, along with any accompanying educational materials. In parallel, we will be developing and deploying monthly workshops as companion resources. For example, we currently offer a monthly “Orange for Beginners”, a programming-free introduction to this software on the basics of visualization and machine learning, which will be incorporated into the DSL-TEACH framework. The final product of DSL-TEACH will be a series of online learning modules along with in-person workshops, offered at varying learning levels. After a full year of implementation, we will conduct a program evaluation to assess its effectiveness.

Conclusion

We have developed a framework that covers the key components of data science along with a prototype of learning modules and assessments to illustrate its utility. DSL-TEACH provides healthcare professionals with the necessary data science conceptual skills to develop intuition and rationale for seamlessly interfacing with tools, such as CDSS, to improve patient care. DSL-TEACH is not a certified program, but rather a community resource covering concepts necessary for training healthcare providers in data science literacy in the context of healthcare decision making.

References

Evaluating EHR Data Availability for Cancer Pain Research
Christina Radloff, BSN, RN1, Jia-Wen Guo, PhD, RN1, Katherine Sward, PhD, RN1
1University of Utah, Salt Lake City, Utah, USA

Abstract

This study identified the availability of cancer pain research related data from the Veterans Affairs (VA) electronic health record (EHR). Two VA nurses independently reviewed 68 concepts and 50 scales identified in cancer pain literature and recorded where and how it is documented in the EHR. Approximately, 80% of the seen concepts are documented in unstructured free text. Techniques such as natural language processing are required to use EHR data for cancer pain research.

Introduction

Pain is one of the most distressing symptoms for cancer patients. The multifaceted, individualized, and dynamic nature of cancer pain necessitates a personalized care approach that considers the complexity and trajectory of cancer pain. The existing longitudinal patient data in the EHR may contain useful information for cancer pain research. However, it is unclear what EHR data are available to advance personalized cancer pain care research.

Methods

In our previous study, we used the Dynamic Symptoms Model to guide the identification of concepts describing cancer pain research. In that study, 68 concepts and 50 scales or measures were identified from literature and classified into 10 domains: including demographic, physiologic, psychological, social, spiritual, environmental, cancer pain intervention, overarching consequences, cancer pain symptoms, and cancer pain-related comorbidities. In the current study, we examined the availability of these cancer pain research related data in the VA EHR documentation systems, including Computerized Patient Record System (CPRS), Bar-Code Medication Administration (BCMA), and Essentris. Two VA nurses, one from the intensive care unit and the other from the general ward, independently reviewed all 68 concepts and 50 scales or measures. During the review, they recorded whether they have seen or documented the concept or scale. The datatype (e.g., structured or unstructured) and the documentation system (e.g., CPRS, BCMA, Essentris) would be further identified if the concept or scale was found in the EHR. Inter-rater reliability (IRR) was calculated to examine the degree of agreement between two nurses.

Results

The IRR results were almost perfect and moderate agreement for scales and concepts, respectively; discrepancies were resolved through discussion. Out of 68 cancer pain research related concepts, 89.7% (n=61) were documented or seen in the VA EHR. Of those 61, 80.3% (n=49) were represented as unstructured free-text. All 61 concepts are stored in CPRS, 8 are in BCMA, and 6 were in Essentris. The 7 concepts not found in the VA documentation systems were types of cancer pain interventions (i.e., cognitive behavior therapy, low-level laser therapy); quality of life; self-efficacy, barriers, and knowledge related to pain management; and social connection. Of the 50 scales or measures, only 8 were seen in the VA EHR: 6 for assessing pain, one for anxiety, and one for cognitive disturbance.

Discussion

This study provides information for researchers interested in using EHR data to study cancer pain. CPRS contains more patient data for cancer pain research than other VA documentation systems because it is the main system for documenting and storing patient records. The majority of the concepts were seen embedded in unstructured, free-text documentation. The concepts that were not seen within VA documentation systems may be found documented using terminology other than the stated concept name. For example, the social support concept was not seen in the VA documentation systems as ‘social isolation’. Social isolation may be documented using surrogate language. For example, ‘the patient lives in a remote area’ or ‘the patient is estranged from their family’.

Conclusion

Extracting EHR data for cancer pain research is not simple since many concepts are found within the unstructured, free-text documentation. Techniques such as natural language processing are required to extract the concept information from free-text. Also, some concepts are not clearly articulated within EHR documentation. Therefore, surrogate concepts and additional variables related to those concepts not seen in VA documentation systems. This will need to be explored in further research.
Addressing the Pressing Home Care Coordination Challenges and Opportunities: A Literature Review

Mohammad Ishtiaque Rahman, BSc, and Güneş Koru, PhD
University of Maryland, Baltimore County, Baltimore, MD

Introduction
Care coordination is pivotal for home health agencies (HHAs) as their participation in value-based purchasing (VBP) model of Centers for Medicare and Medicaid Services (CMS) is becoming mandatory by 2022.1 HHAs can ensure cost effective quality services to their clients which meet the criteria to be reimbursed through better care coordination. Analysis and synthesis of findings from previously published studies can help organizational initiatives aiming to achieve efficiency and effectiveness. Hence, this study has (i) identified the top challenges and opportunities to better coordination of home care services and (ii) derived initiative recommendations for HHAs to improve care coordination.

Methods
A literature review was performed, because such a review provides opportunities to analyze and synthesize information and results from a number of studies. To implement a systematic approach, the PRISMA guideline has been adopted.2 Keywords were developed according to the standard definition of care-coordination stated by AHRQ: Coordination, IT, collaboration, competency, discharge process, and home care giver.3 The relevant publications were selected based on a set of inclusion criteria: being published in US, written in English, only involving home health agencies and employees. Following this method, 7,423 papers were identified for preliminary screening searching IEEE, Pubmed and SAGE databases. A set of 30 papers were shortlisted for subsequent qualitative synthesis by examining the titles and abstracts.

Results
The conducted study reveals the followings as top challenges: (i) Power imbalance of communication among HHAs, physicians and hospitals, e.g., clinicians are not always confident while communicating with a physician and asking questions regarding important patient information. (ii) A lack of integration among EHR systems, as well as an inability to capture contextual information, e.g., EHR systems used in HHAs often do not interact with hospital EHRs and very few EHRs integrate NLP features to capture relevant contextual information. (iii) Alarm fatigue is also becoming relevant, e.g., hospitals and physicians often receive too many requests at once to handle. (iv) Often the HHAs receive lower priority and attention from the hospitals and physicians as they share information, e.g., in many cases HHAs received their requested patient history and medication data too late or did not receive at all. Sometimes, visiting clinicians are perceived as guests by the home care givers, i.e., home caregivers are not perceiving these clinicians as serious medical personnel to share patient information. Despite of these challenges, there are also opportunities which can bolster HHA coordination initiatives. Sharing photos and videos containing rich information using smart phones can increase patient-provider engagement. (i) Physicians are busy but generally willing to communicate, e.g., when contacted in person, physicians are sharing information. (ii) Clinicians enjoy and would likely to continue working in HHAs, reducing coordination problems regarding stuff overturn. (iii) More community resources are now available to the home care clinicians and care givers aiding care coordination, e.g., lists of shared durable medical equipment are publicly available.

Conclusion
The recommended directions emerging from this study are: Using online video streaming training, centralized portal to share and retrieve patient information reducing alarm fatigue, and meaningful use of social networks helping patients and care givers to find information resources. Further research is required addressing issues such as, EHR integration among HHAs and hospitals to improve collaborative effort. NLP research on health care domain will also be useful as right now EHR is not equipped enough to capture relevant contextual information.

References
Predicting Empiric Antibiotic Coverage based on Patient Factors

Protiva Rahman, BS, Erinn M. Hade, PhD, Courtney Dewart, MS, Yuan Gao, MS, Courtney Hebert, MD,MS
The Ohio State University, Columbus, OH

Introduction
While antibiograms provide pathogen-specific coverage details, physicians make syndrome specific decisions for empiric treatment. There has been preliminary work in modeling coverage for individual antibiotics using patient factors such as demographics, recent antibiotic exposure, etc. However, combining outputs of multiple models to recommend an ‘ideal’ one remains an open problem\(^1\). Selecting a high probability threshold can lead to prescribing broad-spectrum antibiotics, increasing antibiotic resistance in the community. In this work, we look at methods to identify the narrowest effective antibiotic for each patient.

Method
Our data consisted of adult patients admitted to an inpatient unit of an academic medical center from 11/11/11 to 6/30/16 with a urinary tract infection (UTI) and a positive culture in first 48 hours of admission (N=6,366, train = 5,093, test = 1,273). To rank antibiotics we used a mixed-methods approach of a survey followed by in-person discussion with the university’s antibiotic stewardship group\(^2\). We used the mean breadth score (on a scale of 100) from the 9 survey-responses as an antibiotic’s breadth. To model coverage, we used five penalized logistic regressions, one for each antibiotic (Cefazolin, Ceftriaxone, Ciprofloxacin, Cefepime, and Zosyn). As a first step in determining a recommended antibiotic, we selected the antibiotic with the lowest breadth score and an estimated probability above a specific threshold. For each patient, we further calculated the ‘ideal’ antibiotic as the narrowest (lowest breadth) antibiotic to which a patient was sensitive, based on lab results. We then trained hierarchical feed-forward neural networks with two fully-connected hidden layers to directly identify the ideal antibiotic.

Results
We evaluated the recommended antibiotic on two measures: the breadth score and the actual coverage (number of infections that would be successfully treated if the recommended antibiotic was prescribed). Figure 1 shows the actual coverage and breadth trade-off for varied thresholds of penalized and neural-network models in our test dataset. We also show results for physician administered and ideal antibiotic. A threshold of 0.65 covers 65% of infections at avg. breadth score of 60. Other thresholding schemes (0.65 for Cefazolin and Ceftriaxone, 0.7 for Cipro, 0.8 for Cefepime, otherwise Zosyn), result in covering 70% of UTI cases with breadth score of 62. Neural networks perform comparable to penalized models. As expected, increasing coverage, requires increasing breadth.

![Image of Figure 1](https://example.com/figure1.png)

*Figure 1. Coverage and breadth trade-off for different thresholds.*

Conclusion
We compare the trade-offs of different thresholding schemes when ranking antibiotics by combining predictions from five models. Our models are able to cover the same number of infections as the physician with a 5 point decrease in breadth, and they can cover 10% more infections with the same breadth as physicians. However, they are still far from the ideal scenario of covering 85% of infections at 43.5% breadth.

References:
Approaches to Pilot Testing of Electronic Exchange of Infectious Diseases Case Report Data to Public Health Disease Surveillance System in Minnesota

Sripriya Rajamani, MBBS, PhD, MPH, FAMIA1,2,3; Ann Kayser, MPH1; Hanna Chakoian, BA1; Sarah Boneske, MPH1; Emily Emerson, BS1; Sarah Solarz, MPH1

1Minnesota Department of Health, St. Paul, MN; 2School of Nursing, 3Institute for Health Informatics, University of Minnesota, Minneapolis, MN;

Introduction

Public health agencies collect surveillance data on certain communicable diseases to assess disease trends, characterize affected population and prioritize disease control strategies. In Minnesota (MN), reportable diseases are determined by the Minnesota rules1 and communicable disease reporting is centralized. Reporting sources submit data to the MN Department of Health (MDH) and surveillance is supported by the Minnesota Electronic Disease Surveillance System (MEDSS)2. An estimated 71,995 cases were reported in 2017, most of which occurred through phone, fax, paper-based disease report cards and web-based data entry. With exponential growth of electronic health records (EHRs) in the clinical sector, there is increasing momentum to move towards electronic case reporting (eCR) to public health propelled by national initiatives. Currently, MDH is exploring a multi-modal approach to eCR to support stakeholders.

Methods

Driven by various factors ranging from national/local initiatives and the need to respond to time-dependent system changes by clinical data submitters, more than one approach was needed. Three different options (Table 1) for electronic exchange of case report data was considered. Preliminary steps/planning underway for the first one (Digital Bridge), with pilot test planned for later in the year. Last two options (HIE and eCR Lite) are under implementation.

Table 1: Approaches to Support Electronic Case Reporting

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Digital Bridge3                Nationally scalable, multi-jurisdictional approach with automated generation and transmission of case reports from EHRs to public health</td>
</tr>
<tr>
<td>2</td>
<td>Health Information Exchange   Data exchange supported by state-certified Health Information Exchange (HIE) service providers in Minnesota (HIE Oversight Program)4</td>
</tr>
<tr>
<td>3</td>
<td>eCR Lite2 (MN specific)        Customized MN approach that builds on national guidance along with additional data elements</td>
</tr>
</tbody>
</table>

Results

Rules for six nationally identified infectious diseases are currently being authored as part of Reportable Condition Knowledge Management System (RCKMS) which is integral to the national Digital Bridge initiative (option #1). A health system that is a MN-HIE certified entity (option #2) started reporting in early March 2019. This system decided to utilize eCR Lite approach (option #3) to submit their data. Regular data feeds (~8-12 cases/day) with relevant data (~84 out of 101 data elements in eCR Lite specs) extracted from their EHR through a reporting module are being submitted to MEDSS. A mid-sized provider clinic system was also able to successfully test eCR Lite (option #3) and is switching to that reporting method shortly. A customized data exchange method (part automated, part manual) was implemented to support receipt of data through option #3 in the short term.

Conclusions

MEDSS plans to move towards stream-lined and standards-based methods to support electronic data exchange of reportable case information. There is a need to share case studies and lessons learned to promote interoperability.

References

Clinical Trials Process Framework to Optimize Clinical Research Operations

Pallavi Ranade-Kharkar, MS, PhD1, Benjamin Chisum, BS1, Kathryn G. Kuttler, PhD1
1Intermountain Healthcare, Salt Lake City, UT

Introduction
Clinical research operations – from initial hypothesis to research publication and closeout – is complex, time-consuming, and resource-intensive. The operational processes are further complicated by regulatory requirements and involvement of multiple stakeholders, from siloed teams, some of whom are external to the organization. To improve efficiencies associated with administrative and other operational processes, it is crucial to define standardized workflows across departments and create a framework of measurement based on milestones defined at the appropriate level of granularity. Doing so, provides an opportunity to improve processes, provide transparency into workflows, track operational metrics, assist with resource estimation and optimum utilization, and provide data-driven rationale to leadership for requesting additional resources.

Method
We conducted semi-structured interviews with 37 stakeholders representing seven functional teams in the research pipeline (e.g. Institutional Review Board (IRB), contracts, billing compliance) and major research departments (e.g. Cardiovascular, Oncology, Women and Newborns) to understand their current workflows. We then created future-state workflows that identified critical milestones and their sub-processes (we called them actions) that: 1) indicate completion of an atomic activity (e.g. completion of the contract intake process); and/or 2) handoff responsibility from one team to another (e.g. initial billing plan essential documents uploaded where responsibility is passed from the research department to billing compliance); and/or 3) involve historically time-consuming processes (e.g. budget negotiation). We further defined sets of milestones for each of the types of studies, namely: sponsored, investigator-initiated, and data or sample study. We conducted three rounds of refinement to the milestones and actions. A detailed review of the milestones was completed with the stakeholders to ensure a balance between the need to record data points for tracking metrics and the data entry overhead for users. These milestones were implemented in a third-party Clinical Trial Management System (CTMS) that currently includes 502 studies. Stakeholders and users of the CTMS were interviewed post-implementation to evaluate ease of use of the process framework.

Results and Discussion
The clinical trial process framework is comprised of 12 milestones (Table 1). Users reported that completing data entry for the milestones and actions in the framework was easy to complete and involved minimal time overhead.

Table 1. Clinical research milestones and actions
<table>
<thead>
<tr>
<th>Milestone name</th>
<th>Milestone description</th>
<th># of actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feasibility</td>
<td>Outlines the steps in accessing study feasibility</td>
<td>4</td>
</tr>
<tr>
<td>Pre-award</td>
<td>Guides pre-award grant creation and submission</td>
<td>2</td>
</tr>
<tr>
<td>Billing Plan</td>
<td>Outlines steps for completing billing plan and Medicare coverage analysis</td>
<td>8</td>
</tr>
<tr>
<td>IRB approval</td>
<td>Tracks IRB submission and approval</td>
<td>2</td>
</tr>
<tr>
<td>Budget</td>
<td>Tracks steps involved in negotiating a budget with a sponsor</td>
<td>7</td>
</tr>
<tr>
<td>Contract</td>
<td>Outlines steps involved in creating/managing a research contract</td>
<td>5</td>
</tr>
<tr>
<td>Study validation</td>
<td>Outlines validation checks prior to activating a study for enrollment</td>
<td>5</td>
</tr>
<tr>
<td>Study startup</td>
<td>Outlines processes for accounting and administrative costs</td>
<td>3</td>
</tr>
<tr>
<td>Study closeout</td>
<td>Outlines steps for closing out a study</td>
<td>10</td>
</tr>
<tr>
<td>Quality assurance</td>
<td>Outlines processes to ensure quality parameters are met for a study</td>
<td>7</td>
</tr>
<tr>
<td>Amendment</td>
<td>Tracks steps for managing contract, billing plan, or IRB amendment</td>
<td>8</td>
</tr>
<tr>
<td>Physician contract</td>
<td>Outlines processes involved in completing a physician contract</td>
<td>4</td>
</tr>
</tbody>
</table>

Conclusion
A process framework based on standardized workflows was designed to capture critical milestones in clinical research operations. It appears to be feasible to track accurate and meaningful clinical research metrics and optimize clinical research operations using the proposed process framework.

References
eHealth in Military Primary Care: How Consumers Engage Electronically with the Military Health System

Stephanie J. Raps, MSN-BC, PhD Student
Graduate School of Nursing, Uniformed Services University, Bethesda, MD

Introduction

Primary care teams are becoming increasingly aware of the need to go beyond contemporary clinical approaches and include the use of patient engagement electronic health (eHealth) technologies to enhance patient-centered care. This awareness is emerging from, the effects of a major generational shift surfacing in primary care clinics. A recent study evaluating primary care providers found a split between consumers ages 18-29 ‘not’ having a primary care provider (45%), compared to older consumers 30-49 (28%), 50-64 (18%) and over 65 (12%)1. Consumers are increasingly seeking more convenient healthcare solutions – electronic health (eHealth), virtual visits, retail clinics, and urgent care clinics. However, despite a high consumer interest and expansion of eHealth tools, widespread adoption remains low2,3. The TriCare Online Patient Portal and Secure Messaging application is one of the highest utilized eHealth tools in the MHS and a focus of adoption efforts4. The MHS faces similar eHealth adoption issues while striving to increase engagement and maintain patient-centered, coordinated, and longitudinal relationships. The MHS reports 70.5 million annual visits across 375 outpatient clinics, but little is known about what portion of these visits are achieved by eHealth modes and how consumers interact electronically with the MHS. Our study investigated the utilization patterns of eHealth consumers in primary care within the MHS to describe the characteristics and patterns of use by high utilizers of eHealth and low utilizers of eHealth.

Methods

A retrospective review of the TriCare Online Patient Portal/Secure Messaging utilization metrics was completed. Two dependent variables were assessed: website login times a year and application used (Appointments, Blue Button, Service Separation, Secure Messaging, Nurse Advice Line, and RX Refill). Five independent variables were examined: Military Treatment Facility (MTF), TriCare region, sex, consumer type (Active Duty, retired, dependent) and age.

Preliminary Results

Preliminary data identify Army MHS consumers use the application more than other services. Additionally, viewing and downloading of health record data are the most used MHS eHealth functions. Non-Active Duty females between the ages of 25-35 are the highest users. (results pending April 2019)

Conclusion

These results are expected to have a vital positive impact that provides the missing knowledge to develop innovative education and technological designs that enable patients and providers to communicate satisfactorily and effectively.

References

Visualizations to Communicate Risk in Patient Reported Outcomes
Meghan Reading Turcchio, PhD, MPH, RN\textsuperscript{1}, Lisa V. Grossman, M.Phil\textsuperscript{2}, Annie C. Myers, MA\textsuperscript{1}, Ruth M. Masterson Creber, PhD, MSc, RN\textsuperscript{1}

\textsuperscript{1}Department of Healthcare Policy and Research, Division of Health Informatics, Weill Cornell Medicine, New York, NY; \textsuperscript{2}Department of Biomedical Informatics, College of Physicians and Surgeons, Columbia University, New York, NY

Introduction
Patient-reported outcomes (PROs) are gaining popularity as a means of supporting patient self-management. Recent research has begun to identify the smallest changes in PRO scores that are clinically meaningful (i.e., minimally important differences, or MIDs) for different patient populations\textsuperscript{1}. However few tools exist to communicate MIDs to patients with different health literacy, graph literacy, and numeracy levels. Communication of MIDs is critical for patients to understand and act on PRO scores. Visualizations of personal health data have the potential to convey risk to a diverse range of patients, but have predominantly been tested with data that have well-established boundaries for action (i.e., lab values)\textsuperscript{2}. The purpose of this study was to assess patients’ risk perception and behavioral intention in response to visualizations that display MIDs in PRO scores over time.

Methods
We recruited a purposive sample of hospitalized heart failure (HF) patients from an inpatient cardiac unit at an urban academic medical center to participate in individual interviews. Participants were English-speaking adults diagnosed with HF. A professional designer developed four visualizations (text only, text plus icon, text plus number line, and text plus line graph) depicting side-by-side comparisons showing worsening MIDs of PRO scores over time. We showed participants the colored, paper-based visualizations during their inpatient visit. Using the participants’ preferred visualization of the four, we evaluated subjective risk perception\textsuperscript{3,4}, objective risk perception\textsuperscript{5}, and behavioral intention. We then collected demographic and literacy-related surveys. Data were analyzed using SAS 9.4.

Results
Participants (n=40) had an average age of 61.3 years (±12.5) and were 22% female, 52% White, and 38% Latino. They reported having low education (high school or less: 45%), financial resources (43%), health literacy (55%), graph literacy (91%), and subjective numeracy (mean score 12.2 ±4.7). Most also had mild/moderate cognitive impairment (88%). Participants’ favorite visualizations were icons (31%) and number lines (41%). Subjective risk perception was generally high; 70% reported it was very/extremely likely their HF was getting worse and 54% reported they were very/extremely worried about their HF getting worse. However objective risk perception was mixed; 67% accurately recalled the gist of the visualization but only 10% accurately recalled the PRO scores verbatim. Regarding behavioral intention, most (82%) were very/extremely likely to act based on the visualization. The majority of these participants reported their next step would be to contact a healthcare provider or go directly to a hospital (84%), while fewer reported they would attempt to self-manage the symptom or search the Internet for health information (16%). The 8 participants who were unlikely to act reported it was due to their perception of low risk.

Conclusion
Visualizations of MIDs of PRO scores over time communicated risk to the majority of patients, who in turn reported being more likely to act. Patient decision-support may bolster patient skills and efficacy to self-manage, and minimize burden on healthcare providers, when patients perceive risk from PRO scores.

This work was supported by NINR of the NIH under Award Number R00NR016275 (PI: Masterson Creber).

References
Technology-Mediated Adverse Events in Primary and Community Care

Chantelle Recksky, RN, MSN, PhD(c)1, Lorraine Blackburn, RN, MSN2, Allison Muniak, MSc2 Kathy Rush, RN, PhD3, Maura MacPhee, RN, PhD1, Leanne M. Currie RN, PhD1
1University of British Columbia-Vancouver, British Columbia, Canada;
2Vancouver Coastal Health, Vancouver, British Columbia, Canada;
3University of British Columbia-Okanagan, Kelowna, British Columbia, Canada

Introduction
A technology-mediated adverse event (TMAE) is a patient safety event with a potentially negative outcome (i.e., harm) in which health information technology played a role. The incidence of TMAE is increasing in parallel with high uptake of technology in healthcare. Most TMAE research has been conducted in acute care settings. A study of general practitioners in Australia found up to 6 TMAEs per practice over 19 months (with 6.6% having potential for harm). The purpose of this study was to examine rates of TMAEs in primary and community care settings.

Methods
This cross-sectional study was carried out in a Canadian healthcare region providing care for over 1 million people in urban and rural settings. Primary care is provided via an MD or NP. Community and public health care is provided at a clinic or the patient’s home. A web-based tool for voluntarily reporting patient safety events has been used since 2009. Reports include a narrative description, categorization of type (e.g., medication), and degree of harm. In 2015, a screening question derived from Magrabi et al.,’ was added: ‘Was the computer system a factor in the event?’ We analyzed reports from primary and community care from November 1, 2016 to October 31, 2018.

Results
Over 2 years, 105 events were tagged as “yes” to the screening question. These reports were investigated and verified by local supervisors. Of these, 19 (18%) had harm attributed – minor (n=13), moderate (n=5), and severe (n=1). Medication events represented 65% of reported events (n=68). Among the different areas of care, public health had the most reported events (n=38, 36%), yet 92% (n=35) of these events indicated no harm (e.g., vaccines administered outside of vaccine schedule). In contrast, 6% (n=6) of events were reported in home health, and half of events (n=3) in this setting indicated harm (e.g., missed care due to information loss in care transition).

Discussion
To the authors’ knowledge, this is the first study that reports on frequency of TMAEs in community care. Bias associated with self-reporting is a limitation. Qualitative data are being gathered and analyzed as another phase of this work. Healthcare leaders need to be aware of the potential for harm related to advancing technologies in practice, and encourage recognition and reporting of TMAE. These findings will guide the health region to establish a process to identify, analyze, and address TMAE in non-acute settings.

References
HomeSHARE: Implementing Multi-Site Smart Technology Infrastructure

Blaine Reeder, PhD\textsuperscript{1}, Haley Molchan, BS\textsuperscript{2}, Eric Gutierrez, MSPH\textsuperscript{1}, Ewelina Pena, BS\textsuperscript{1}, Kelly Caine, PhD\textsuperscript{3}, PhD, George Demiris, PhD\textsuperscript{4}, Katie Siek, PhD\textsuperscript{2}, Kay Connelly, PhD\textsuperscript{2}

\textsuperscript{1}University of Colorado, Aurora, CO; \textsuperscript{2}Indiana University, Bloomington, IN; \textsuperscript{3}Clemson University, Clemson, SC; \textsuperscript{4}University of Pennsylvania, Philadelphia, PA

Introduction

Smart home, wearable, and mobile sensor technologies are increasingly ubiquitous and have potential to enable previously unavailable opportunities to support research in daily living settings. However, investigators who conduct technology research in home environments face challenges of recruitment, generalizability of results, reusability and obsolescence of technology, data and algorithm transparency from proprietary devices, and sustainability of research efforts beyond funded study periods. The Home-based Smart Health Applications across Research Environments (HomeSHARE) project is a multi-site smart technology infrastructure initiative (NSF Award #s: 1629202, 1629468, 1625451, and 1629437) implementing a distributed testbed to integrate and evaluate a diversity of devices in participant homes and serve the needs of multiple research communities including those of informatics, gerontology, computer science, nursing, and public health. HomeSHARE sites are located at University of Colorado (Colorado), Indiana University (Indiana), University of Pennsylvania (UPenn) and Clemson University (Clemson).

The current HomeSHARE implementation relies on 1) a smart home system: sensors (motion and contact sensors) installed in the home that transmit activity data to a hub via wi-fi then to a project database via the internet; 2) system servers: servers located in a data center with a research console for project configuration, sensor data export, and remote hub management; 3) smart watch: wearable activity monitors that transmit activity data via bluetooth to a tablet computer (then vendor and HomeSHARE databases); and 4) participant portal: a tablet computer with touchscreen that allows researchers to collect electronic self-reported data from survey instruments and conduct remote video interviews. Participant self-reported survey data are transmitted to a REDCap database.

Methods

A cross-site infrastructure demonstration study using an identical mobility monitoring protocol\textsuperscript{1} in the homes of independent living participants 65+ years of age began in October 2018. Sensor and self-reported data are being integrated and analyzed for correlations. Technology implementation factors and feature needs are being documented for infrastructure improvement.

Results

Eighteen participants (n=14 Indiana; n=4 Colorado) are enrolled. Enrollment is ongoing with planned enrollment of 15 participants at both sites. Demonstration study results have informed requirements that will result in transitioning HomeSHARE to a new internet-of-things platform. Figure 1 shows a future state HomeSHARE architecture with backend features for improved integrated data reporting and remote technology management through a single research console.

Conclusion

We have successfully deployed a multi-site smart technology infrastructure designed for multiple research communities. An enhanced infrastructure study will scale HomeSHARE at Colorado, Indiana, UPenn, and Clemson (n=100 homes). Participant sensor data visualizations and implementation findings will be presented at AMIA 2019.

Acknowledgements

Thank you to Jeffrey Kaye, MD and OrcaTech for technical partnership and Marjorie Skubic, PhD for consultation.

References

Initial Phase of Developing a Model to Predict Unplanned Readmissions in the Australian Private Healthcare Sector before they Occur: the Case of Hysterectomy

Daniel Reischel, MS1, Isabella Eigner, MS1, Freimut Bodendorf1, Jonathan L. Schaffer2, MD, MBA, Nilmini Wickramasinghe, PhD, MBA3
1FAU, Nuremberg, Germany; Cleveland Clinic, Cleveland, Ohio, USA; 3Swinburne University of Technology & Epworth HealthCare, Melbourne, Victoria, Country, Australia

Introduction: In the Australian private healthcare sector there is a growing pressure to reduce unplanned hospital readmissions, i.e., readmissions related to the initial indexed visit that were unexpected within 28 days of discharge, which are considered a quality of care indicator as well as a major cost driver1. The Australian Institute for Health and Welfare (AIHW) focuses on reducing unplanned readmissions post specific surgical procedures. Hysterectomy has been identified as the cause of the 2nd highest unplanned readmission within these procedures2; yet no model to predict unplanned readmissions following a hysterectomy exists. Australia has one of the highest rates of hysterectomy (27,586 procedures a year) in benign diseases as compared to other OECD countries3. Hence, this study focusses on developing a suitable predictive model which has lessons for other cases and contexts of unplanned readmissions.

Methodology: A double de-identified data set with 800 patient data records over 10 years at a private not-for profit tertiary healthcare facility is examined. To this data set, processes suggested by Shmueli and Koppius4 are employed to build a suitable prediction model. Clinical data, coupled with clinical insights extracted from focus groups were used. Inclusion criteria for assessing unplanned readmissions included readmission (based on DRG code) related to indexed initial visit within 28 days of discharge from indexed visit.

Discussion: Especially with imbalanced data, novel ensemble approaches promise a better performance in predicting the minority class. These approaches combine resampling techniques with either boosting or bagging. The most promising ensembles being RUSBoost or underbagging. To have a comprehensive benchmark also SMOTEBoost and Overbagging are added to determine the best performing approach in predicting unplanned hysterectomy readmissions. Since all ensemble methods need base classifiers, this study compares how decision tree (DT), artificial neural net (ANN) and support vector machine (SVM) classifiers perform in combination with these novel ensemble methods and their components. To identify the best model, the AUC-Score is used as performance measure.

Table 1 Performance comparison for different modelling approaches

<table>
<thead>
<tr>
<th>Base model</th>
<th>RUS Sampled</th>
<th>SMOTE Sampled</th>
<th>Bagging</th>
<th>Under bagging</th>
<th>Over bagging</th>
<th>AdaBoost</th>
<th>RUS Boost</th>
<th>SMOTE Boost</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>0.71</td>
<td>0.93</td>
<td>0.88</td>
<td>0.58</td>
<td>0.93</td>
<td>0.90</td>
<td>0.66</td>
<td>0.94</td>
</tr>
<tr>
<td>ANN</td>
<td>0.50</td>
<td>0.48</td>
<td>0.52</td>
<td>0.50</td>
<td>0.53</td>
<td>0.56</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SVM</td>
<td>0.51</td>
<td>0.51</td>
<td>0.55</td>
<td>0.50</td>
<td>0.52</td>
<td>0.58</td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>

The best performing models are combinations of DT with random undersampled data or ensemble methods implementing random undersampling. The best performing model is RUSBoost with an AUC-Score of 0.94. The base ensemble methods bagging and Adaboost achieve worse scores than the novel approaches. Since Adaboost and its derivations RUSBoost and SMOTEBoost cause overfitting, underbagging should be preferred.

Conclusion: We built and compared multiple predictive models that indicate far reaching implications for theory and practice. We observe that novel ensemble approaches increase the predictive performance but note that these improvements. As this approach has far reaching implications it is therefore important to examine data sets from other healthcare organizations to confirm the obtained results which will form the basis of our future work.

References:
Promoting Trust, Standards, and Real-World Applications for Patient-Centered Clinical Decision Support

Joshua E. Richardson, PhD, MS, MLIS¹, Laura Marcial, PhD¹, Blackford Middleton, MD, MPH, MSc²; Kensaku Kawamoto, MD, PhD, MHS³; Jerome A. Osheroff, MD⁴; Beth Lasater, MSPH¹; Barry H. Blumenfeld, MD, MS¹
¹RTI, International, Research Triangle Park, NC; ²Averpixa, Inc., Chicago, IL; ³University of Utah, Salt Lake City, UT; ⁴TMIT Consulting LLC, Naples, FL

Abstract
Clinical decision support (CDS) increasingly encompasses content to support patients and clinicians alike, such as decision aids for shared decision-making. Actions are therefore necessary to attain reliable and meaningful content for “patient-centered CDS” (PCCDS). The Patient-Centered CDS Learning Network led three parallel efforts to promote PCCDS: addressing issues of trust in PCCDS content; recommending standards for PCCDS; and developing an “action plan” for implementing PCCDS to manage chronic pain and opioids.

Introduction
PCCDS is CDS that directly involves or addresses patient-specific needs in partnership with caregivers. The Patient-Centered CDS Learning Network – funded by the Agency for Healthcare Quality and Research (AHRQ, Cooperative Agreement #5U18HS024849-03) – engaged numerous stakeholders from 2018 to 2019 to deliver products that inform vendors, policy-makers, patients (and more) about ways to structure evidence from patient-centered outcomes research and make it more trustable, standardized, and implementable within compelling real-world use cases.

Methods
The Learning Network formed three working groups (WGs) comprised of an array of stakeholders including patients, providers, policy-makers, content and electronic health record (EHR) vendors, and more. The WGs developed charters to govern themselves and define the outputs for promoting PCCDS-related actions. The WGs routinely met from February 2018 – January 2019 to address issues of trust, standardization, and implementation; and developed white papers that were community vetted, edited, and approved for release by AHRQ representatives.

Results
The three WGs identified a variety of challenges and developed recommendations that provide an overarching agenda for promoting PCCDS. The recommendations include the areas of policies, content authoring, implementation, and measuring impact. Table 1 summarizes the outputs from each WG.

Table 1. Working Group (WG) Results, Impacts, and Next Steps

<table>
<thead>
<tr>
<th>Trust Framework WG</th>
<th>Technical Framework WG</th>
<th>Opioid Action Plan WG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>33 recommendations organized by nine “trust attributes” including competency, compliance, consistency, patient-centeredness, and transparency</td>
<td>Six core areas of challenges, facilitators, and recommended actions along with examples in the pain management domain</td>
</tr>
<tr>
<td>Impact</td>
<td>Recommendations are being integrated into CDS Connect, an AHRQ-funded CDS repository and authoring platform</td>
<td>Disseminating output to stakeholders (e.g. HL7) that supports PCCDS development and implementation</td>
</tr>
<tr>
<td>Next Steps</td>
<td>Promote the recommendations as part of Learning Network efforts and integrate into learning health system scholarship</td>
<td>The Learning Network implements a set of core recommendations for piloting a standards-based patient-facing app in 2019</td>
</tr>
</tbody>
</table>

Discussion
The Learning Network developed a rich content base in three areas that are critical to the promotion of PCCDS: content trustability, technical standardization, and multi-stakeholder collaboration. The results provide elemental pieces to a broad ecosystem of policies and technological solutions that will be essential for sustaining PCCDS, and CDS more broadly. We believe our efforts contribute to a growing body of work pertaining to decision support solutions in clinical care as well as within learning health systems. The Learning Network will build upon these efforts and continue promoting PCCDS as a key area of development in the informatics sphere.
Racial Disparity in Clinical Decision Support Use for Acute Upper Respiratory Infection

Safiya Richardson, MD MPH1, Thomas McGinn MD MPH1, Simon Jones PhD2, Joseph Palmisano MD3, Sara Chokshi PhD2, Catherine Dinh-Le2, Rebecca Mishuris MD MA3, Linda Park PhD4, Paul D. Smith MD4, Ainsley Huffman5, Sundas Khan MD1, Rachel Hess MD5, David Feldstein MD MA2

1Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY; 2New York University School of Medicine, New York, NY; 3Boston Medical Center, Boston, MA; 4University of Wisconsin School of Medicine and Public Health, Madison, WI; 5University of Utah School of Medicine, Salt Lake City, UT

Introduction: Clinical decision support (CDS) improves quality of care by improving diagnosis, treatment, and preventative care. CDS can standardize care and eliminate health disparities by providing unbiased care recommendations. CDS has been shown to eliminate racial disparities in child abuse screening and venous thromboembolism prophylaxis—in both cases White patients were receiving inferior care, lower screening and prophylaxis rates. The ability of these tools to reduce disparities depends on unbiased provider use. The objective of this study was to evaluate provider use of CDS by patient race in the context of a large randomized controlled trial.

Methods: This study was a part of a controlled trial, randomized by site, including 11 primary care practices at a large academic healthcare system in Wisconsin. Between October 2015 and June 2018 providers were presented with CDS to assist in estimating pre-test probability of group A strep pharyngitis and pneumonia with the goal of reducing unnecessary antibiotic prescriptions for acute upper respiratory infection. In patients with the chief complaint of sore throat providers were presented with a tool built with the Centor Score and in those with a chief complaint of cough or upper respiratory infection they were presented with a tool built with the Heckerling Rule for Pneumonia. Providers who completed the calculator portion of the CDS tool were considered users. Electronic health record data on provider and patient demographics and CDS tool use rates were collected for analysis.

Results: The intervention group included a total of 27,801 patient encounters with tool launch. The patient racial breakdown was: American Indian 0.68%, Asian 2.83%, Black 3.77%, Native Hawaiian and Pacific Islander 0.18%, White 91.57% and Other 0.97%. Overall CDS tool use was 9.4% (CI; 9%-9.7%). CDS use among Black patients was 13.4% (CI; 11.4%-15.5%). Providers were more likely to use the tool with Black patients (OR=1.51 (CI; 1.25-1.81), p <0.001) compared to White patients in a logistic regression model adjusted for age, gender and site clustering. Tool usage rates among the remaining race categories were similar to overall rates. In a model additionally controlled for provider type there was no racial difference in tool use. Antibiotic prescriptions were lower in Black patients (OR=0.84 (CI; 0.17-0.97), p=0.019).

Conclusions: Provider use of CDS in this trial varied by patient race. This may be explained by resident provider’s increased likelihood to use the tool and to see Black patients. Providers were more likely to use the tool with Black patients which could exacerbate a current disparity as White race is associated with higher antibiotic prescription rates for acute upper respiratory infection. Further study is needed to understand this difference in CDS use rates.
Semi-automated clinical data annotation framework using web-based RedCap and Jupyter Notebook Tools

Albert William Riedl, MS; Irene Cortes Puch, MD; Ernest Pierce Stewart, DO; Nicholas R Anderson, PhD; Jason Yeates Adams, MD

1: University of California at Davis. Sacramento, CA

Introduction: Secondary use of electronic medical record (EMR) data has proven more challenging than anticipated but is possible with careful curation. While developing an automated SOFA score annotation pipeline, the need for rapid cycle QA became apparent. The Sequential Organ Failure Assessment (SOFA) score is a tool to measure dysfunction in multiple organ systems, with higher scores associated with a higher mortality. While there are other illness severity scores (eg APACHE), one benefit of SOFA is that the data elements are, in theory, readily accessible. In practice, however, automated SOFA scoring requires integrating repeated measures, temporal data alignment, and handling charting errors. A rapid Quality Assurance (QA) cycle is necessary to increase experimentation velocity, enable iterative development, and provide benchmarks for success. We developed an extensible, web-based framework for consensus development and evaluation of gold standard annotation by interdisciplinary teams of clinicians and data scientists.

Methods: A RedCap form captures manually derived gold standard annotations from clinicians. Computed annotations were generated by a Python script and saved to a SQL database. The gold standard vs computed comparison was implemented as a Jupyter Notebook, a publishing format for reproducible workflows. The Notebook extracted computed annotations via SQL and gold standard annotations via the RedCap API (Figure 1). Annotations were index aligned via study identifier. SOFA organ system component scores were aligned pair wise by source (clinician vs computed) in columns to enable easy identification of discrepancies (Table 1).

Table 1: Index aligned, score pair visual review. Score pairs represent the 6 components of the SOFA score

<table>
<thead>
<tr>
<th></th>
<th>coag</th>
<th>ov</th>
<th>cns</th>
<th>hepatic</th>
<th>renal</th>
<th>respabg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Results: 80 critical care encounters were adjudicated via root cause analysis in interdisciplinary team meetings using this system to target anomalies, accelerating the development cycle and improving trust in computed annotations. The visual feedback helped classify errors as point failures or systemic issues. In our early work, we have found that most discordant scores are caused by errors in the gold standard annotation process itself; reviewers making errors that were not present in the computed scores.

Discussion: Accurate transformation of raw EMR data into meaningful information frequently requires algorithmic processing that must be guided by manually derived clinician gold standards. Gold standard data set development is time consuming and there are no widely adopted tools that facilitate rapid comparisons between clinician annotations and algorithmic output. This framework efficiently identified errors in algorithm logic, and importantly, in the manual chart review process rather than assuming that clinician annotations are inherently accurate. Future work will include adding a discrepancy resolution workflow in the Jupyter Notebook to facilitate consensus adjudication of scoring differences and descriptive statistics to better characterize algorithm accuracy relative to the final gold standard. This approach is extensible to any set of index aligned scoring comparisons.

References
Using Ontology-Driven Decision Support For Hereditary Cancer Risk Assessment

Jordon B. Ritchie,1 Lewis Frey, MS, PhD,1 Jean-Baptiste Lamy, PharmD, PhD,2 Heath Morrison, BS,1 Brandon M. Welch, MS, PhD,1
1Medical University of South Carolina, Charleston, SC,
2LIMICS, Université Paris 13, INSERM UMRS 1142, Sorbonne Universités, 93017 Bobigny, France

Introduction

Identifying patients who should be recommended for cancer genetic counseling and possible screening, relies heavily on family health history (FHx). FHx is comprised of family relationships, health related information for family members such as cancer history and demographic information, and is most often represented as a pedigree chart. This type of data represents a challenging data model to analyze. Additionally there are very specific evidence based guidelines, such as hereditary cancer guidelines from the American College of Medical Genetics\(^1\) (ACMG), that detail scenarios where patients should be recommended for counseling based on their FHx. These evidence based guidelines are complicated semantic rules that cancer genetic counselors (CGC) have to evaluate by hand. There are hundreds of rules in the ACMG guideline for hereditary cancer alone not to mention many additional guidelines. Currently CGCs are expected to evaluate hand-drawn pedigrees against hundreds of complicated semantic rules. This likely leads to failure to recommend patients, and family members of patients, for genetic counseling in cases where genetic counseling is needed.

Approach

Our proposed solution for storing collected FHx and applying evidence based semantic rules to determine whether patients or their family members should be recommended for genetic counseling hinges on Ontology Oriented Programming (OOP) and Semantic Reasoning. Ontologies have been around for a long time and are especially good at representing data that have important relationships that need to be preserved in the data. This is true for FHx. Ontologies allow the data to be stored in a graph model that depends on triples that define two objects (individuals) in the graph joined by their relationship (properties). This easily allows us to represent a family pedigree and their associated health information in the ontology. We can also define rules in our ontology that will guide semantic reasoners in reclassifying individuals of a given class into subclasses based on their properties that define their relationship to other classes in the ontology. Semantic reasoners evaluate triples and rule definitions to perform this classification process. The recent development of the Owlready\(^2\) API gives us programmatic access to our ontology and comes with built in access to the Hermit reasoner\(^3\) which is the industry standard reasoner that is also built into Protegé ontology editor\(^4\). This allows us to declare the ontological classes, properties, individuals, and rules and perform reasoning all in memory without using an editor such as Protegé.

The ability to programmatically define our ontology, store data in the ontology, and perform semantic reasoning allows us to create a service that consumes FHx and produces recommendations using ReST technology.

References


Rebecca L. Rivera, PhD, MPH, CPH\textsuperscript{1}, Saurabh Rahurkar, PhD\textsuperscript{2}, Brian E. Dixon, MPA, PhD, FACMI, FHMSS\textsuperscript{1,3}, Joshua R. Vest, PhD, MPH\textsuperscript{1,3}, Nir Menachemi, PhD, MS\textsuperscript{1,3}, Ying Zhang, PhD\textsuperscript{1,4}, Jason T. Schaffer, MD\textsuperscript{4}, Julia Adler-Milstein, PhD\textsuperscript{5}, FACMI, Titus K. Schleyer, DMD, PhD, FACMI, FAMIA\textsuperscript{3,4}

\textsuperscript{1}Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN; \textsuperscript{2}CATALYST Center, the Ohio State University Wexner Medical Center, Columbus, OH; \textsuperscript{3}Regenstrief Institute, Inc., Indianapolis, IN; \textsuperscript{4}Indiana University School of Medicine, Indianapolis, IN; \textsuperscript{5}Department of Medicine, University of California, San Francisco, CA

Introduction

Due to the paucity of health information exchange (HIE) evaluations employing causal study designs, demonstrations of positive effects of HIE implementations have largely remained theoretical. Our project presents the development of an enhanced methodology using a causal study design to evaluate the effect of HIE on healthcare processes and outcomes. We will apply this novel methodology to evaluating an innovative intervention that directly integrates HIE into electronic health records (EHR).

Objective

Our objective was to develop a novel and comprehensive method to evaluate CareView, an app integrating HIE information directly into Cerner\textsuperscript{1}, on user perceptions, and healthcare processes and outcome measures.

Methods

Our pragmatic evaluation applies a causal study design integrating objective measurements of HIE use with its qualitative context through surveys and focus groups to understand CareView’s impact. We will roll out CareView to 15 Indiana University Health emergency departments using a stepped-wedge cluster non-randomized study design that allows for each cluster to serve as a control. Clinicians and patients will be followed over time to connect the impact of HIE use to outcomes on a patient-by-patient basis, not only at a population-level as in previous studies. This mixed-methods evaluation is guided by a framework combining two models previously validated in the context of health technology interventions: the Unified Theory of Acceptance and Use of Technology II (UTAUT), and the DeLone and McLean Information Success Model. Data from clinician-user logs and patient records will be used to quantify HIE use. In addition, barriers to and motivators for using HIE will be uncovered through qualitative measures, providing context around specific use cases. We will relate these measures to care processes and outcomes including readmissions, and repeat tests and orders.

Policy Implications

HIE innovations to improve efforts toward meaningful integration of patient data must be guided by high-quality evidence provided by rigorous evaluations such as described for this project. This type of evaluation is necessary to provide evidence for continued financial and legislative support of interoperability by federal and state agencies, as well as other stakeholders.

Conclusion

Our proposed evaluation helps provide a practical and pragmatic roadmap to determine causal effects of HIE/EHR integration on meaningful HIE use, and its impact on improving healthcare processes, quality and patient outcomes.

Application of Semantic Web Technology to regulations

Mitra Rocca, Dipl.-Inform.Med¹, Rashed Hasan, PhD¹
¹US Food and Drug Administration, Silver Spring, MD, USA

Introduction:
The U.S. Food and Drug Administration’s (FDA) regulatory activities are based on federal law and Title 21 of the Code of Federal Regulations (21 CFR). CFR is an annual codification of the general and permanent rules published by the executive departments and agencies of the federal government of the United States. The entities regulated by FDA, including pharmaceutical, biotechnology, medical device, dietary supplement, food, cosmetics and tobacco companies, must follow the regulations in 21 CFR in all their operations under FDA jurisdiction. Depending on the product being regulated, these operations may include, but are not limited to, product manufacture and controls, nonclinical and clinical studies, submissions to the agency for human research or product marketing, labeling, and post-market safety reporting. By applying Semantic Web Technology (SWT), we developed a semantic model that enabled us to establish a linked resource for Part 11 of 21 CFR. This resource may be useful in clinical protocol design and phases of clinical development. The model can be extended to other parts of 21 CFR (and other regulations) to enable greater efficiency to the by providing a linked source of regulations.

Methods: For our project, we used AlchemyAPI [1]. AlchemyAPI uses Natural Language Processing (NLP) algorithms to extract metadata from the 21 CFR documents. We ran the application programming interface (API) over 21 CFR Part 11, which contains 10 sections. After discussion with domain experts, we considered only those phrases that were not longer than two words. We obtained 146 keyphrases of one or two words. Many of these, although highly ranked concepts, had nothing to do with the text, while seemingly obvious concepts in part of Alchemy’s library were not returned. Therefore, domain experts reviewed the 146 keyphrases and validated 86 of them, while classifying the remainder as false positives. Furthermore, to enable the system to automatically explore the relevant new datasets from the external sources, we connected the 21 CFR Part 11 keyphrases to the Linked Open Data (LOD) cloud. The proposed solution provides a visualization that shows how keyphrases are connected with each part and subpart of the regulation. This visualization provides a useful alternative representing relations.

Results: The system organizes the data into three layers: (1) User interface (UI), (2) Middleware, and (3) data source. The developed UI allows users to perform keyphrases search via a web page. In the Middleware, Solr does the analyses on the user input data and creates an index into the system. All of the Part 11 of 21 CFR section files were indexed in the Apache Solr indexing system[2], which is the foundation of the 21 CFR web-based query system. For better search functionality, we added NGram functionality (either NGramFilter-Factory or EdgeNGramFilterFactory) in the Solr configuration file. When the user performs full text searches, user-entered keyphrases will match any prefix of indexed text, rather than requiring the full word to be entered.

Conclusions: We have conducted an analysis of 21 CFR regulations Part 11. By applying Apache Solr indexing mechanisms and having subject matter experts validate keyphrases, we have demonstrated the potential to develop a usable search application. In addition, we presented an approach for publishing 21 CFR Part 11 keyphrases into the LOD Cloud and developed a search platform using Solr that supports users when searching documents by keywords. Future efforts may include expanding this work to other parts of 21 CFR regulations. Additionally, this approach may be applied to other FDA regulations and guidances, policy document from other health authorities, and the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

References


Acknowledgements: Funding support received from the FDA/CDER/OTS and FDA ORISE Program

Disclaimer: The views expressed are those of the authors and do not necessarily represent the views of the US FDA or the US Government.
Mobilizing Million Hearts: Using Workflow Analysis In The Optimization Of Clinical Decision Support Tools

Ronald E. Romero¹, Lucy C. Stein, MS OTR/L¹, Laura C. Schubel¹, Christine Laccay, MHA, Mark Townsend², MD, A. Zachary Hettinger MD¹, Kevin Chaney, MGS³, Kristen E. Miller; DrPH, CPPS¹

¹National Center for Human Factors in Healthcare, MedStar Health, Washington DC, USA; ²Centra Health, Lynchburg VA, USA; ³Office of the National Coordinator, Washington DC, USA

Introduction

The goal of our research is to develop and pilot a clinical decision support (CDS) tool that aims to optimize the current state of atherosclerotic cardiovascular disease (ASCVD) risk calculators. CDS has been shown to improve provider performance, with automated and workflow integrated CDS performing better than user initiated, nonintegrated CDS. The research team mapped clinician stakeholder’s workflow to leverage the ASCVD risk assessment tool and SMART (Substitutable Medical Applications, Reusable Technologies) on FHIR (Fast Healthcare Interoperability Resources) technology to integrate cardiovascular risk assessment into the electronic health record (EHR).

Methods

Participants were recruited for stakeholder interviews and clinical observations across a large Mid-Atlantic healthcare system to ensure diversity of care providers. Provider feedback, use cases, and observed tasks were used to better understand the desired features and functions of our novel health information technology ASCVD risk tool. By analyzing clinician workflow and needs, optimized functions were mapped to the appropriate technological capabilities needed to implement new functionality into a workflow embedded app.

Results

A total of 17 clinician interviews were conducted with an interdisciplinary group representative of the ASCVD surveillance process: cardiologists (n=6), primary care physicians (n=7), and care navigators (n=4). Additionally, 30 hours of clinical observations were conducted. Because of the physician workflow analysis, a critical objective of our optimization process for the ASCVD risk tool has become the auto population of ASCVD risk scores and guideline-based treatment recommendations to help decrease clinician burden.

Conclusion

The results of our research revealed implications for both technical design and development of the optimized ASCVD risk tool. Similar EHR tools used to aid recognition of abnormal blood pressure levels have seen an increase in recognition of elevated BP despite minimal app usage, leading researchers to believe workflow integration would maximize the usefulness of EHR CDS tools. The research team is expecting an increase in usage of the ASCVD risk calculator as a patient education tool during preventative cardiovascular risk discussion once fully integrated. The technical team continues to explore the use of SMART, SMART on FHIR, and CDS Hooks as well as native EHR integration methods (e.g., CCL, Smart Templates) to increase functionality.

Table 1. Sample functionality of optimized ASCVD risk calculator mapped to corresponding technical requirements.

<table>
<thead>
<tr>
<th>Function</th>
<th>Feature</th>
<th>Technical Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculates risk with minimal workload</td>
<td>Auto populate values into the calculator</td>
<td>FHIR and CCL call</td>
</tr>
<tr>
<td>Ensure ASCVD risk is up-to-date</td>
<td>Alert to missing/outdated information</td>
<td>FHIR and UX</td>
</tr>
<tr>
<td>Recalculate score</td>
<td>Refresh risk for new values</td>
<td>CDS Hooks</td>
</tr>
<tr>
<td>Evaluate risk score trends</td>
<td>Display risk scores (current and previous)</td>
<td>Mpage with FHIR component</td>
</tr>
</tbody>
</table>

PGxKnow: A Pharmacogenomics Educational Application in Augmented Reality

Don Roosan, PharmD, PhD; Western University of Health Sciences, College of Pharmacy, Pomona, CA, USA; Huy Truong, BS, PharmD Candidate; Keck Graduate Institute, School of Pharmacy, Claremont, CA, USA; Jay Chok, PhD, Keck Graduate Institute, Riggs School of Applied Life Sciences, Claremont, CA, USA; Moom Roosan, PharmD, PhD; Chapman University, School of Pharmacy, Irvine, CA, USA.

Background: Pharmacogenomics has the potential to reduce adverse drug events and improve health by individualizing medication therapy. Despite such potential, studies have shown that poor understanding of pharmacogenomics concepts by patients may lead to misinterpretation of the test results and decreased utilization. Simple videos or presentation materials were not very intuitive or interactive to grasp the complex genomics concepts for patients. We propose PGxKnow, an interactive pharmacogenomics augmented reality-based educational platform that can deliver information to patients in an engaging manner. The objective of the application is to educate patients on complex pharmacogenomics concepts to improve comprehension so that the patients recognize the lifelong implication of pharmacogenomics in medication therapy.

Overview of methodology: We conducted a human-factors-based methodology, cognitive work analysis (CWA), to create specific educational components that improve comprehension of pharmacogenomics concepts. Two pharmacogenomic trained pharmacists identified the important pharmacogenomics concepts. These concepts were further stratified into different frames for interactive features using 3D models for cells, nucleus, chromosomes, and proteins. Sequence of frames were arranged for better clarity through several iterations of feedback from the expert faculties. We used UNITY and Microsoft Visual Studio to transform the concepts into the interactive augmented reality PGxKnow application. Two pharmacists and two research assistants provided their initial feedback about the application on a quick usability study. We developed the application in a Microsoft Hololens device.

Results: We successfully created PGxKnow, an interactive application that shows frame by frame how mutations in genes may cause protein malformations with suboptimal function. Thus, these proteins contribute to increased or decreased drug metabolism leading to increased adverse events or treatment failure. The different visualizations and interactive features allow patients to have a broader understanding of how mutations from DNA can have a detrimental effect on drug metabolism. We incorporated feedback from the quick usability from the four participants and reiterated the design of the application.

Conclusion: In this study, we have successfully created an educational application, PGxKnow, for educating patients on pharmacogenomics concepts using augmented reality. In future, we plan to conduct usability testing of our application and assess how effective this application is in improving users’ understanding of pharmacogenomics concepts.

Acknowledgement: This work was supported by an internal grant from Keck Graduate Institute and Western University of Health Sciences, College of Pharmacy.

References:

Retrospective analysis of electronic health records reveals the ineffectiveness of a prescribing intervention

Lauren M. Rost, BA¹, Erik S. Wright, PhD¹
¹Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania

Abstract
Antibiotic stewardship programs (ASPs) facilitate the appropriate prescribing of antibiotics within clinical care. These programs are recently mandatory in U.S. hospitals, and are upheld to certain standards, like utilizing health data for quality improvement. However, due to the nascent stage of these regulations, the field has not standardized the adoption of EHR data to inform and validate prescription interventions. We retrospectively analyzed outcomes associated with an implemented intervention to evaluate its effectiveness and inform future interventions.

Introduction
Antibiotic stewardship programs (ASPs) regulate unnecessary and inappropriate prescribing of antibiotics, and have been mandatory for U.S. hospitals since 2017¹. Part of this mandate includes the utilization of health data for the improvement of ASPs. However, due to the nascent stage of these regulations, most hospitals have yet to evaluate the effectiveness of their ASP². Electronic health record (EHR) mining presents an opportunity to standardize methods for evaluating and improving stewardship interventions.

Objective
In 2009, the University of Pittsburgh Medical Center’s (UPMC’s) ASP issued a stewardship intervention to switch from treating community-acquired pneumonia (CAP) patients with ceftriaxone plus azithromycin to ampicillin-sulbactam plus azithromycin, with the goal of decreasing the prevalence of C. difficile infections. We sought to determine clinician adherence and patient outcomes of this intervention.

Methods
We performed a retrospective analysis on EHR data from a cohort of CAP patients, identified through 70 ICD-9/10 codes, at 17 UPMC hospitals from 2004 to 2017. We determined adherence by assessing the distribution of antibiotics prescribed for the cohort. We calculated Cox proportional hazards for the effect of treatment on length of stay, in-hospital mortality, 30-day readmission rates, antibiotic duration, and time-to-therapy.

Results
Preliminary analyses have shown that there was low adherence to the intervention in 15 of the 17 UPMC hospitals, with the two hospitals that demonstrated adherence having more interactions with the ASP. We also found that the intervention did not have a substantial effect on the prevalence of C. difficile infections.

Conclusion
These findings emphasize the potential value of big data approaches to inform the impact of stewardship interventions. In advancing our analyses further, we seek to predict alternative successful treatment options using machine learning models.

References
Combining Forces for MRI Brain: Natural Language Processing of Radiology Reports with Structured Documentation

Justin F. Rousseau, MD, MMSc¹, Juan Diego Rodriguez, MS², Jared Abrams, PhD², Alex Liu, PhD², R. Nick Bryan, MD, PhD¹

¹University of Texas at Austin, Dell Medical School, Austin, TX; ²University of Texas at Austin, Applied Research Laboratories, Austin, TX

Introduction

There are tradeoffs with structured and unstructured approaches to healthcare documentation. Structured documentation offers opportunities for more accurate, complete, and computable data, while free-text allows physicians to expressively describe their thought processes. Galileo CDS® is a diagnostic clinical decision support (CDS) tool with an interface for radiologists to choose structured values for 12 imaging features for brain MRIs (e.g., FLAIR signal, contrast enhancement). The features are used to generate a probabilistic differential diagnosis and a standardized report¹². Free-text reports and structured data entry through the CDS tool follow two distinct paths of synthesis of images, clinical data, and medical knowledge, to produce an interpretation. We sought to understand how these paths might merge by converting radiology reports into the structured feature values via NLP for the CDS tool.

Methods

106 brain MRI radiology reports were assigned values for all CDS features by two non-clinician annotators who had received introductory training from an expert neuroradiologist. The reports were used to predict three inputs to the CDS tool using NLP methods: "increase" for Flair and T2, and "Yes" for Contrast Enhancement (CE). Inter-annotator agreement was measured with Cohen’s kappa coefficient (Flair: 0.47, T2: 0.61, CE: 0.53). All text was then preprocessed with Natural Language Toolkit (NLTK)³, including tokenizing, converting to lower case and removing stopwords. We experimented with both stemming and using term-frequency-inverse document frequency (tf-idf) weighting. Due to the limited amount of data, we focused on three text classifiers known to work well in low-data regime: logistic regression, multinomial naive Bayes, and complement naive Bayes, all implemented in scikit-learn⁴. We applied stratified 10-fold cross-validation 10 times and evaluated the methods with precision, recall, and F1 scores.

Results

Complement naive Bayes had both higher average F1 scores and less variance than multinomial naive Bayes and logistic regression. The best results were achieved using complement naive Bayes with stemming and tf-idf weighting in the preprocessing pipeline. Average F1 scores for the FLAIR, Contrast Enhancement, and T2 signal were 74.9%, 77.7% and 82.3%, respectively.

Discussion

Our methods of predicting the value of features based on NLP of radiology reports shows promise that feature values can be predicted using automated NLP techniques where with more data, predictions can be further improved.

Conclusion

Future work is needed to take further steps to extract additional features from clinical notes and evaluate methods of comparing the CDS tool with features extracted from the radiology reports. Future work is needed to compare NLP-extracted data to radiologist-entered data in the CDS tool and to compare mentions of diagnoses in the interpretations to the predicted diagnoses from the CDS tool and to true diagnoses vetted in the EHR.

References

Comparing De-Implementation Strategies: Education vs Removing Orders from Order Sets

Elise L. Ruan, MD¹, Joshua M. Pevnick, MD²

¹Albert Einstein College of Medicine, Bronx, NY, ²Cedars-Sinai, Los Angeles, CA

Introduction: Traditionally, order sets have been used to address issues of underuse by reminding providers of necessary orders. However, recent evidence suggests that order sets may contribute to overuse. In an era where overuse concerns loom large, we must evaluate the role of order sets in inadvertently encouraging overuse, and the potential role of order set modification in reducing overuse.

Methods: In our health system, we suspected overuse of creatine kinase (CK) testing. American College of Cardiology guidelines recommend against routine CK testing for asymptomatic patients on statins. Because CK ordering frequently came from an ambulatory order set used by one cardiology group, we removed the CK order from this order set. We then compared CK order frequency before and after order removal. We also compared order removal with an educational intervention applied to this cardiology group (Group A), as well as educational interventions applied to two other cardiology groups (Groups B and C) not using this order set.

All three groups received educational interventions, which included group and one-on-one talks about appropriate indications for CK testing (Table 1). We retrospectively defined a study period including one year before and after the date of order removal (9/21/17). For each provider, we extracted data on how many CK tests were ordered and whether orders came from the order set. To control for provider volume, we identified all ambulatory visits to each group. Finally, two investigators, including a cardiologist informatician, reviewed a random sample of 20 charts from encounters resulting in a CK order to evaluate order appropriateness.

Results: During the study period, there were 77,567 visits to all three groups and 6,906 CK orders. 403 of CK orders were placed through the order set. 5% of providers accounted for 66% of orders in the pre-intervention period. Without adjusting for provider volume, order set modification was the only intervention associated with a significant decrease in the order/visit ratio for providers (p=0.03). When adjusted, results showed that order set modification in Group A and the two educational interventions in Group C achieved significance. Order removal on average resulted in a 50% relative decrease (p=0.01) while the first educational intervention in Group C resulted in a 38% relative decrease (p=0.02). The second educational intervention resulted in a 90% relative decrease (p=0.02).

Table 1: Effect of Interventions on Provider Order/Visit Ratios

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
<td>Before</td>
</tr>
<tr>
<td>Order Removal (9/21/2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visits</td>
<td>9,356</td>
<td>6,177</td>
<td>N/A</td>
</tr>
<tr>
<td>CK Orders</td>
<td>723</td>
<td>268</td>
<td>N/A</td>
</tr>
<tr>
<td>Order/Visit Ratio</td>
<td>0.08</td>
<td>0.04</td>
<td>N/A</td>
</tr>
<tr>
<td>p-value (95% CI)</td>
<td>p = 0.009 (-0.06, -0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educational Intervention #1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention Date</td>
<td>5/15/2017</td>
<td>10/2017*</td>
<td>10/17/2017</td>
</tr>
<tr>
<td>Visits</td>
<td>6,177</td>
<td>3,589</td>
<td>4,900</td>
</tr>
<tr>
<td>CK Orders</td>
<td>268</td>
<td>46</td>
<td>477</td>
</tr>
<tr>
<td>Order/Visit Ratio</td>
<td>0.04</td>
<td>0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>p-value (95% CI)</td>
<td>p = 0.165(-0.08, 0.01)</td>
<td>p = 0.050(-0.08, 0.00)</td>
<td>p = 0.016 (-0.13, -0.02)</td>
</tr>
<tr>
<td>Educational Intervention #2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention Date</td>
<td>N/A</td>
<td>N/A</td>
<td>4/2/2018</td>
</tr>
<tr>
<td>Visits</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CK Orders</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order/Visit Ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Educational meetings were conducted with individual providers over the month; thus, no specific date is listed.

Discussion: Both order set modification and educational interventions were associated with decreases in test ordering. Advantages and drawbacks for each type should be considered when trying to change ordering behavior. Order set modification can theoretically be done without any additional involvement from providers who are often already limited in their availability. In cases where a small number of providers account for a large portion of inappropriate orders, a targeted educational intervention may be a simpler solution, but would not be as scalable as employing clinical decision support systems and order set modification. Similar to how both education and order sets are currently widely used to ensure important orders are not missed, there is potential for the same approaches to be applied to de-implementation and reducing overuse.
Leveraging temporal patterns in physiological variables for prediction of critical events in single ventricle infants

Victor M. Ruiz, PhD1, Jorge Galvez, MD1,2, Michael Goldsmith, MD1, Maryam Naim, MD1,2, Alejandro Lopez-Magallon, MD3, Ricardo Munoz, MD3, Fuchiang (Rich) Tsui, PhD1,2

1Children’s Hospital of Philadelphia, Philadelphia, PA
2University of Pennsylvania, Philadelphia, PA
3Children’s National Medical Center, Washington, DC

Introduction: Congenital heart defects (CHD) affect approximately 40,000 births annually in the United States. Single-ventricle (SV) defects are among the most complex CHDs, and infants in this population are at elevated risk of unexpected deterioration in the peri-operative period before stage-2 palliation.1

There is a scarcity of models for prediction of critical events (CEs) suggesting cardiopulmonary decompensation in SV infants, i.e., extracorporeal membrane oxygenation (ECMO), emergent endotracheal intubation (EEI), and cardiopulmonary resuscitation (CPR). State-of-the-art models2–4 achieved high performance (0.91 AUC) in the hour leading to CEs, but suffer from low early prediction performance, require significant technology investments from institutions, or ignore temporal trends in physiological variables. In this study, we extracted temporal information from longitudinal data of SV ICU admissions, and trained static and dynamic machine-learning classifiers for early prediction of CEs in SV infants. We hypothesized that leveraging longitudinal changes in routinely-collected physiological data would help predict the onset of CEs with up to eight hours of anticipation.

Methods: Dataset: We retrieved data from 271 SV infants who were admitted to the ICU before six months of age and before undergoing stage-2 palliation at the Children’s Hospital of Philadelphia between January 1, 2015 and September 30, 2018. Variables in the dataset included vital signs (e.g., heart rate, respiratory rate, blood pressure, oxygen saturation), and laboratory tests (e.g., creatinine, bicarbonate ion, base excess.) From these 271 infants, we identified 164 CEs as cases, and randomly selected 164 periods of non-eventful ICU admissions as controls. Patients with multiple CEs were considered as separate cases provided that they presented at least eight hours apart.

Longitudinal patterns: We extracted two sets of temporal features from longitudinal data. First, trend-summary features that summarize changes for each variable, e.g., slope of last two values, difference between apex and last values. Second, frequent temporal patterns (FTP), which capture trajectories of multi-variate temporal abstractions derived from longitudinal data, e.g., heart rate is high before oxygen saturation starts decreasing.

Modeling approach: We trained static classifiers, i.e., naïve Bayes, support-vector machine, decision tree, and random forest models from trend-summary and FTP features. We also trained long short-term memory models from raw time-series data. We trained and evaluated models at five prediction horizons, i.e., one, two, four, six, and eight hours before the onset of CEs. We measured predictive performance with the area under the ROC curve (AUC).

Results: Random forest classifiers trained with FTP features achieved the highest predictive performance, with AUCs of 0.86 (95% CI: 0.84-0.87), 0.84 (0.83-0.86), 0.84 (0.83-0.85), 0.86 (0.85-0.87), and 0.85 (0.84-0.86) at prediction horizons of one, two, four, six, and eight hours before CEs, respectively. At one hour and eight hours before CEs, the best model had sensitivities of 0.6 and 0.59, and PPVs of 0.87 and 0.86 at the 0.9 specificity level, respectively.

Conclusion: In this study, we leveraged the temporal information in longitudinal, objective data from SV infants admitted to the ICU to meet the need of accurate and early prediction of CEs. We achieved this by using data that can be extracted automatically from EHR systems, and that is routinely collected. Early prediction of CEs may enable clinicians to target interventions aimed at reducing morbidity, mortality, and healthcare costs associated with SV defects.

Correspondence: Rich Tsui, PhD, Department of Anesthesiology and Critical Care Medicine (email: tsuif@chop.edu)

References

Reasons for Non-participation in an Inpatient Acute Health Portal Randomized Controlled Trial Among Hospitalized Patients
Beatriz Ryan-Lora, MPH1, Fernanda Polubriaginof, PhD1, Irma Alarcon1, Niurka Suero-Tejeda, Lisa Grossman2, Ruth Masterson-Creber, PhD3, David K. Vawdrey, PhD1
1NewYork-Presbyterian Hospital, New York, NY, 2 Columbia University Medical Center, New York, NY, 3Weill Cornell Medicine, New York, NY

Introduction and Background: Various studies have extensively examined the barriers and adherence to patient health portals in the outpatient setting. However, not many have examined barriers to use of patient health portals during hospitalization. We examined non-participant responses from a pool of participants who declined to be part of a randomized controlled trial (RCT) on Understanding Patient Information Needs During Hospitalization.1 The study aimed to randomize participants into three study arms: 1) usual care, 2) tablet with general access to the Internet and 3) tablet with an inpatient portal. We explained that the portal provided a variety of information ranging from vitals, basic medication lists, lab orders and results, and scheduled procedures such as CT scans or Ultrasounds.

Methods: Two bilingual (English/Spanish) research assistants recruited patients admitted to two separate step down units at Columbia University Medical Center between March 2014 and May 2017. Recruitment methods for the larger RCT are recorded in previously published work.1 Patients who declined to participate were asked if they approved to provide a recorded answer as to why they refused to participate. Research assistants used a script approved by the Institute Review Board. Participants who agreed to be interviewed were informed that their answers would be anonymous. They were asked to answer a 5-item survey that included open-ended questions about age, gender, race and reasons for non-participating. Answers were recorded on paper surveys and transcribed into excel format for further analysis. The open-ended questions were categorized into eight main categories and 22 subcategories of reasons for refusal. Descriptive statistics were calculated using SPPS.

Results: From the pool of recruited patients approached to participate in the RCT, 195 refused to participate but agreed to be interviewed about the reasons for their refusal. Of those interviewed, 62.1% were male, 59.5% were white, and 80% chose English as a preferred language. The median age of the respondents was 66-75 (28.2%) followed by 56-65 (26.7%). Categories of patient refusals are listed in Table 2. From our analysis we noted that the main reasons why patients refused to participate were due to General Denial (27.2%) because of pending discharge (12.3%) or generally not being interested in participating (12.3%). The second main category of refusal was listed under Vitality or Health reasons (25.6%), with 12.8% of non-participating patients expressing “worry regarding health outcomes.” The third category for refusal was Negative Feelings about Technology (17.4%), the main subcategories being “hesitation to use or adopt technology” (16.15%), “reluctance to use of technology” (5.13%) and “dislike of technology” (4.10%), with a small fraction stating they were “too old to learn” (3 participants) and preference of “face to face interaction” (1 participant).

Discussion: Previous studies have shown that admitted patients have a desire to access their medical information2,3 however, adoption and adherence to patient health portal use in the inpatient setting remains a challenge for many healthcare institutions. Despite many assertions that better-informed patients are more equipped to manage their medical care, some of our findings indicate that barriers remain in adoption of this type of technology by admitted patients due to pre-occupation with health outcomes and negative perceptions of technology.

References:
A Novel Classification of Inherited Retinal Dystrophies

Hamideh Sabbaghi, MS1, Sina Madani, MD, PhD2, Hamid Ahmadieh, MD1, Narsis Daftarian, MD1, Fatemeh Suri, PhD1, Abbas Sheikhtaheri, PhD3, Farid Khorrami, PhD4, Proshat Saviz, BS3

1Shahid Beheshti University of Medical Sciences, Tehran, Iran; 2Vanderbilt University Medical Center, Nashville, TN, USA; 3Iran University of Medical Sciences, Tehran, Iran; 4Hormozgan University of Medical Sciences, Bandar Abbas, Iran

Introduction

Inherited retinal dystrophy (IRD) as one of the rare ocular diseases encompasses different types of hereditary retinal degenerations appearing by involvement of the different layers of the posterior ocular segment1-2. In respect of the importance of the classification systems regarding the recorded history of biologic and natural sciences3, there is no classification regarding the IRD diagnoses. Therefore, we aimed to present a novel classification in this regard.

Methods

For developing this classification, an expert group of two academic board certified retina specialists, one academic clinical and molecular geneticist, and five medical informatics specialists was formed. Initially, a comprehensive review was conducted on the reference text books in the field of ophthalmology focusing on the vitreous and retina, afterwards a primary classification of IRD diagnoses was developed. In the next step, a comprehensive search was conducted on different international classification systems including Unified Medical Language System (UMLS), Online Mendelian Inheritance in Man (OMIM), International Classification of Diseases (ICD-10 & 11), Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) and Orphanet Rare Disease Ontology. The primary classification was modified and finalized based on the consensus of working group members at several focus group meetings. Eventually, an organized hierarchy of IRD diagnoses was developed in the Protégé software. We classified our results as no matched, partially matched or completely matched. The IRD content was considered as term and concept matched if the term was completely in accordance with our search results. Further evaluation regarding concept matching was performed for no-term-matched IRD contents. Completely concept matched was also defined in cases with synonyms. In SNOMED-CT, we considered post-coordination if any synonym was not found. In fact, if there were possibilities to develop a concept post-coordinately, we considered that concept as “partially concept matched”. IRD contents with no possibility of pre- and post-coordination were considered as “completely no matched”.

Results

Based on the literature review, the primary classification including 55 IRD diagnoses had been provided in six sections. Afterwards, it was expanded to 1020 IRD diagnoses presenting by different international classification systems. As a result of the several panel group discussion, the final list containing 368 IRD diagnoses with consideration of both phenotyping and genotyping identifications in the eight sections was developed. Regarding the comparison between the five international classification systems, it was found that UMLS had the greatest coverage of 92.4%, while the lowest content coverage (24.7%) was obtained for SNOMED-CT. In general, all investigated systems had 12.8% overlap for IRD diagnoses. Furthermore, we found that 0.09% of diagnoses cannot be considered as IRD although they were categorized by other systems. All preferred names, synonyms or other alternative names and codes of each retinal dystrophy entity were extracted and documented, as well.

Conclusion

This specialized classification that is dedicated for different types of IRD diagnoses may be applicable in the ophthalmic practices and it can be considered as a preliminary step to develop an ontology for IRD concepts.
Workflow Pipeline for Medical Image Data Curation and Sharing

Sylvester Sakilay MS¹, Danne C. Elbers MS², Scott T. Doyle PhD¹, Luis E. Selva PhD²,
Brett R. Johnson PhD¹,²,³, Nhan V. Do MD²,³,⁴, Peter L. Elkin MD¹,³,⁵
¹University at Buffalo, Buffalo, NY; ²MAVERIC, VA Boston Healthcare System, Boston, MA; ³BD-STEP, VHA Washington DC; ⁴Boston University School of Medicine; ⁵VA Buffalo Health System

Introduction

Many advancements in medicine have been made possible by the collaboration between researchers and organizations via data sharing. Medical imaging is a critical source of data in research and patient treatment. Sharing the medical data with the external organizations has several important aspects to be taken care of. Most importantly, obtaining the patient consent to release their data and removing the PII (Personal Identifiable Information)¹. We have experienced several issues in handling the image data along the way and have gone through cycles of evaluation of various possibilities. This experience has helped us develop a pipeline that handles this process from the point of acquisition of imaging data to the point of delivery in a more efficient manner. The solution consists of an image de-identification tool and tools developed in python for filtering and QA (Quality Assurance). This work is being done as a part of PODR (Precision Oncology Data Repository) project, MAVERIC VA. Our idea is the method that we adopted would save a lot of time and effort for other researchers and organizations who are working on making their data available to external organizations and researchers.

Methods and Materials

We are using the CTP (Clinical Trial Processor) image de-identification tool from TCIA to remove PII from the DICOM files. Image Filter tool filters the images based on modality and type (original or derived) and queries the DICOM headers to produce manifest and metadata files. The data in these files is used to verify that only consented data is allowing to be shared. Unconsented data is not shared. Also, this step gives control over the type and modality of image data to be shared. After verification, the raw image data is sent to the CTP de-identification tool that scrubs PII data from the images. Once the de-identification process is complete, Final QA process is performed using a python tool. There are times when dose reports are embedded with images. This Final QA process guarantees that embedded reports are identified and removed from the data set prior to sharing with researchers. The de-identified DICOM data after successfully passing through these stringent QA checks is moved to the point of dispatch where a metadata file that contains the information such as Md5Checksum is generated using a python tool. From this point the image data is dispatched to the external organizations.

Conclusion

This pipeline helps sharing the data with the approved external researchers and organization while efficiently addressing the most important considerations like patient consent and PII which is a challenge as the image volume is huge and image data requires additional consideration in terms of format, access, and information extraction.

Acknowledgements

Funding was provided under the VA Cooperative Studies Program Study #2010.

References

1. Samuel Ajjarapu, Frank Meng, Danne C. Elbers, Nhan V. Do, Robert Hall, Karen E. Pierce-Murray, Luis E. Selva, Beth Katcher, Brett R. Johnson, Andrew J. Zimolzak, Corri DeDomenico, Mary Brophy, Louis D. Fiore: Releasing De-Identified Clinical, Imaging, and Genomic Data from the VA to External Repositories for the APOLLO Network. AMIA 2017

Figure 1. Pipeline for Image Data Sharing.
Crowdfunding Medical Care: An Exploratory Comparison of Canada, the United Kingdom, and the United States

Sameh N. Saleh MD¹, Ezimamaka Ajufo MD¹, Richard J. Medford MD¹²
1. Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX, USA
2. Division of Infectious Diseases, UT Southwestern Medical Center, Dallas, TX, USA

Background: Despite major differences in their healthcare systems, medical crowdfunding (CF) is now recognized as an important part of the healthcare funding landscape in Canada (CAN), the United Kingdom (UK), and the United States (US). Systematic comparison of the purpose, characteristics, and determinants of the success of CF campaigns in these countries may help to identify gaps in healthcare provision and potential inequities in the medical CF landscape in each setting.

Methods: GoFundMe (GFM) is the largest for-profit company in the medical crowdfunding arena, raising over $650 million dollars per year for healthcare costs. We conducted a cross-sectional analysis of GoFundMe (GFM) campaigns between February 2018 -2019 in CAN, the UK, and the US. Through web scraping, we directly extracted variables from over 1000 campaigns under the “Medical” subheading of each country’s GFM discovery webpage. We explored descriptive statistics for numerical variables extracted and evaluated the correlation matrix between these variables and amount raised. All monetary values were converted to US currency in dollars. We then performed a manual 2-person review of 10% of campaigns for each country to evaluate demographics, diagnosis, type of treatment, and if the campaign was primarily for treatment costs. Concordance was >90% in all categories.

Results: The vast majority of campaign beneficiaries in the three countries examined were non-black (CAN 99.1% vs. UK 99.1% vs. US 95.2%) adults (CAN 77.7% vs. UK 82.3% vs. US 76.6%). Campaign beneficiaries were more likely to be male in the US (60.5%) and CAN (59.8%), but not in the UK (39.8%). US campaigns set higher mean goals ($88,242; SD $180,758) compared to those in CAN ($31,149; SD $47,041) and the UK ($38,345; SD $87,309) and raised more mean funds ($48,647; SD $26,234) than campaigns in CAN ($17,447; SD $17,057) and the UK ($14,480; SD $28,674). Number of donors, GFM hearts, and fundraising goal were strongly positively correlated with the amount raised in all three countries, each with correlation coefficients >0.5. Facebook shares were strongly positively correlated with amount raised in the UK with a coefficient of 0.65, but only moderately correlated in the US and Canada with values of 0.32 and 0.38 respectively. Length of campaign and number of words or characters in the campaign text were not correlated with amount raised. In all three countries, cancer was by far the most reason for fundraising followed by neurologic conditions and trauma. While routine care was the most common reason for fundraising in the US (CAN 23.4% vs. UK 19.7% vs. US 72.1%), fundraising for experimental (CAN 31.9% vs. UK 46.1% vs. US 18.6%) and alternative therapies (CAN 19.1% vs. UK 11.8% vs. US 7%) was more common in CAN and the UK (Figure 1).

Conclusions: We show that GFM medical CF campaigns in the US set and achieved higher fundraising goals than those in CAN and the UK. CF for routine care was by far the most common reason for fundraising in the US, while campaigns for experimental and alternative care were more common in CAN and the UK. When comparing to the general population in each of the three countries, blacks were less likely to be beneficiaries of medical CF campaigns. Further work is required to determine generalizability of these findings in other CF platforms and to elucidate the factors that underlie these differences.

Figure 1. The graphs above display the type of treatment utilized per campaign for each country. The graph on the left reflects all campaigns and the graph on the right reflects only campaigns that were clearly to primarily fund treatment.
eReferral for Enrollment in Smoking Cessation Programs

Abraham Saraya, MD1, Simita Mishra, PhD1, Pat Folan, DNP1, Nathaniel Brown, MA1, Michael Oppenheim, MD1, Jan Horsky, PhD2
1Population Health Informatics, Northwell, Health, Manhasset, NY
2Center for Research Informatics and Innovation, Northwell Health, Manhasset, NY

Introduction

Tobacco use remains to be a significant preventable contributor to morbidity and mortality in the United States. Referral of patients for counseling and brief personal consultations that clinicians have during office visits are effective interventions to increase cessation rate but long-term, intensive assistance to smokers yields better results.1 Completing a full patient referral and following progress, from enrollment into a cessation program to maintaining engagement over time, is time consuming and challenging for clinicians in the context of ambulatory care visits. We have implemented an electronic intervention (eReferral) to simplify and speed up initial referral to a cessation program and to establish a process to monitor and communicate updates on patient tobacco use. It is integrated into an EHR (Figure 1) and allows clinicians to send referrals to the Center for Tobacco Control (CTC) with two mouse clicks. Program counselors contact the patient by phone within 24-74 hours of the office visit, update the EHR record on the patient’s status (Z72.0) during the ambulatory care visit and follow the patient’s progress (eReferral) (Figure 1).

Methods

We queried the clinical data repository of a large healthcare delivery system for patients with newly documented smoking status (Z72.0) during a 21-month study period following the May 1, 2017 eReferral rollout. We counted the number of providers identifying patients as smokers and the proportion of those who used electronic cessation referral.

Results

There were 949 clinicians who identified 5,898 patients as new, active smokers in the first 21 months after rollout. Of those, 307 (32%) used eReferral for 1,892 smokers. The number of clinicians using the tool remained relatively static for the first five three-month periods after rollout before beginning to steadily increase (Figure 2).

Conclusion

We have found that the number of clinicians actively using eReferral initially plateaued after the rollout and the level of engagement started increasing only about a year and a half later. This change was likely due to our training initiative that looked to make more clinicians at new ambulatory care locations aware of this tool and explained the benefits of decreased workload for them and better care quality for their patients. A more routine and automated level of communication between the referring site and the CTC means that clinicians are always informed whether patients follow a path towards lower tobacco use and effectively closes the assessment-referral-treatment loop. The implication for the rollout process is that intensive knowledge dissemination and training are essential for creating conditions that lead to a larger proportion of clinicians using a quality-improvement electronic tool habitually and who can effectively incorporate it into their workflows and ambulatory practice.

References


Figure 1 eReferral screen in the ambulatory EHR results of their enrollment attempt and include a cessation care plan when a patient has joined the program. The objective of this study was to evaluate the frequency of eReferral use during the initial phase after implementation.

Figure 2 Number of providers and referrals of this tool and explained the benefits of decreased workload for them and better care quality for their patients. A more routine and automated level of communication between the referring site and the CTC means that clinicians are always informed whether patients follow a path towards lower tobacco use and effectively closes the assessment-referral-treatment loop. The implication for the rollout process is that intensive knowledge dissemination and training are essential for creating conditions that lead to a larger proportion of clinicians using a quality-improvement electronic tool habitually and who can effectively incorporate it into their workflows and ambulatory practice.

References

Evaluation of Biomedical Citation Selector

Max Savery, Melanie Huston, MS, James Mork, MSc, Olga Printseva, PhD, DSc, Alastair Rae, PhD, Susan Schmidt, MLS, Dina Demner-Fushman, MD, PhD
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction
Finding and accessing scientific literature in PubMed partially relies on indexing articles for MEDLINE® using Medical Subject Headings (MeSH®). Currently, more than 5,000 journals in MEDLINE are fully indexed, where the majority of articles published by these journals have MeSH indexing assigned. However, for approximately 130 selectively indexed journals, articles are only indexed if, after manual review, they are determined to be relevant to biomedicine and the life sciences. In recent years, the number of these articles has increased substantially; for example, in 2018 about 125,000 were added to PubMed. To reduce the number of articles that must be manually reviewed, we have developed the Biomedical Citation Selector (BmCS), a high recall machine learning system that identifies the articles that require indexing. Here, we manually evaluate BmCS’s automated predictions to better understand its performance in the indexing workflow.

Methods
For this evaluation, we manually reviewed 4,763 articles predicted by BmCS to be out-of-scope for indexing at a 99.5% level of confidence. We chose only articles from chemistry journals because BmCS has been shown to perform more poorly on these articles than it does on those from other disciplines. In a subsequent test, we selected 1,248 articles of any discipline, predicted to be in-scope at a 99.5% level of confidence. To manually review the articles, we designed a Microsoft Access form. The title and abstract of each article was reviewed by two of ten different reviewers, and a final label of in-scope or out-of-scope was assigned after considering both reviewers’ input.

Results
The human reviewers agreed with 4,612 out-of-scope predictions generated by BmCS and disagreed with 151. However, 147 of the misclassified articles were publication types that should always be classified as in-scope, such as comments or errata. Because BmCS will handle these with rules in the future, we removed them from the evaluation. Table 1 reflects this adjustment. The reviewers agreed with all 1,248 articles predicted to be in-scope. Accuracy for the in-scope class was computed to be 100% and for the out-of-scope class, 99.9%. Fleiss’ kappa, used to assess interrater agreement, was .981.

Table 1. Agreement, disagreement and accuracy between human reviewers and BmCS

<table>
<thead>
<tr>
<th></th>
<th>Agreement</th>
<th>Disagreement</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-of-scope</td>
<td>4,612</td>
<td>4</td>
<td>99.9%</td>
</tr>
<tr>
<td>In-scope</td>
<td>1,248</td>
<td>0</td>
<td>100%</td>
</tr>
</tbody>
</table>

Conclusion
In order for BmCS to effectively contribute to the indexing pipeline, it must infrequently label in-scope articles as out-of-scope. This is what we observed: After adjusting the predictions as discussed above, only 0.1% of the system’s out-of-scope predictions were incorrect. These results suggest that the system will effectively reject out-of-scope articles from the indexing pipeline. In addition, the perfect in-scope predictions indicate that we will be able to automatically send articles labeled as in-scope with 99.5% confidence to the human indexers, with relatively few false positives. In general, these results demonstrate that BmCS will reduce the number of articles from selectively indexed journals that require manual review by at least the amount we expect, 54%. Further development of BmCS will focus on exploring the effect of more permissive prediction thresholds, as well as integration into the indexing production environment.

References
Predicting adverse event reports in FDA's Adverse Event Reporting System (FAERS) more likely to be useful for causality assessment
Abhivyakti Sawarkar, MD, MMSc1; Robert Ball, MD, MPH, ScM1
1US Food and Drug Administration, Center for Drug Evaluation and Research, Silver Spring, MD, USA

**Background:** The US Food and Drug Administration (FDA) actively monitors the safety of drugs and therapeutic biologics from multiple data sources including FAERS, a database that contains adverse event reports, medication error reports, and product quality complaints resulting in adverse events, submitted to FDA. With an ever-increasing number of reports (> 2 million in 2018), making the review process more efficient so human safety evaluators can focus attention on assessing reports that contain the most useful information for assessing causality between a drug and an adverse event is an important goal.

**Objective:** To accurately predict the adverse event reports in FAERS more likely to provide information useful for causality assessment by safety evaluators using machine learning and text analytics. We hypothesized that certain subsets of reports (industry reports (IR), direct reports (DR), and literature reports (LR)) would have different linguistic patterns in the report narratives that might be useful to distinguish reports more likely to have information useful for causality assessment.

**Data:** Extending previous work, a set of 925 FAERS reports, classified for causality by FDA safety evaluators using a modified version of the World Health Organization–Uppsala Monitoring Centre (WHO-UMC) criteria for drug causality assessment, were coded as a binary classification task by aggregating the causality categories into two groups: group 1. *Certain*, *Probable*, *Possible* & group 2. *Unlikely*, *Unassessable*. Industry and published literature reports are submitted by manufacturers and direct reports are submitted by consumers and healthcare professionals directly to FDA.

**Methods:** Data was divided into train and test sets maintaining proportions of causality categories. We developed two machine learning models; classification & regression trees and random forest, to predict reports less likely to provide useful information regarding causality. We pre-processed the narratives in FAERS reports and used textual features (text length, term frequency, term frequency-inverse document frequency (TF-IDF), N-grams, singular value decomposition (SVD), and cosine similarity for terms in the report narratives among reports) in the models to see which features might predict reports containing useful information regarding causality. No other features were evaluated. We evaluated the accuracy (proportion of true results among the total number of cases examined) of trained classification models. We leveraged 10-fold cross validation and compared the overall model performance.

**Table 1:** Results: Comparison of accuracy of trained classification models using different features from textual data

<table>
<thead>
<tr>
<th>FAERS Reports (n=925)</th>
<th>Classification &amp; Regression Trees</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Term Frequency</td>
<td>TF-IDF</td>
</tr>
<tr>
<td>Industry Reports (854)</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>Direct Reports (75)</td>
<td>0.70</td>
<td>0.75</td>
</tr>
<tr>
<td>Literature Reports (77)</td>
<td>0.85</td>
<td>0.87</td>
</tr>
</tbody>
</table>

**Conclusions:** Our research using FAERS narratives shows that model accuracy improves when carefully selected linguistic features are applied in machine learning models. The random forest model was more predictive and two features; text length and cosine similarity for terms in literature reports, showed higher predictive power to discriminate between FAERS reports with more and less useful information for causality assessment. Further evaluation of this approach with a larger data set and deep learning techniques might improve performance over current machine learning approaches and enable better identification of discriminating narrative features and better predictive power to identify FAERS reports with more useful information for causality assessment.

**References**


**Acknowledgements:** This project was supported by funding support received from FDA/CDER/Office of Surveillance and Epidemiology and Office of Translational Sciences. **Disclaimer:** The views expressed are those of the authors and do not necessarily represent the views of the US FDA, the NIH, or the US Government.
Identifying Population Subgroups in Neuro-fMRI Using Unsupervised Machine Learning

Jenna M. Schabdach¹ MS, Vincent Schmithorst² PhD, Vince Lee², Rafael Ceschin¹,² PhD, Ashok Panigrahy¹,² MD

¹ Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
² Department of Pediatric Radiology, Children’s Hospital of Pittsburgh of UPMC, Pittsburgh, PA

Resting-state functional magnetic resonance images (rs-fMRIs) are often used to examine and evaluate the functional connectivity of different areas of the brain. Because of the MR protocol required to obtain these images, they are highly susceptible to motion. Researchers and clinicians have developed various behavioral and sedation-based protocols to prevent a patient from moving during image acquisition; however, these protocols fail to completely eliminate all sources of motion in an image. Acquired images must pass through one of many motion correction pipelines to account for the spatial and spin gradient effects of motion on the image contents. The combination of machine learning and image reconstruction techniques during and after image acquisition holds great promise for harmonizing images and recovering motion-corrupted data. However, there is little information about the relationship between unsupervised ML techniques and characteristics of rs-fMR images. We suggest that features indicative of high motion images may also be representative of different groups of subjects, but to obtain enough data to test this theory, we must employ the use of a multi-center study. In a multi-center study, the combination of motion artifacts and other factors such as protocol compliance, image harmonization, and subject populations can introduce sources of bias into the data. In our study, we examine rs-fMRI harmonization and motion in a set of complex congenital heart disease (CHD) case and healthy control adolescent subjects acquired through a multi-center neuroimaging study. We develop a set of features to represent global and local motion in each image. First, unsupervised machine learning techniques are used to ensure that the features extracted from the rs-fMRIs are not representative of other factors characteristic of each site, which confirms the success of the image harmonization technique. The unsupervised machine learning techniques were then used to identify other subgroups within the data set using different subsets of the motion feature set. This study suggests that the use of unsupervised machine learning techniques can not only inform harmonization characteristics of multi-center data, but also identify important features of motion artifacts for identifying subgroups within a patient population.
Improving Discoverability of Educational Resources for Clinical and Translational Scientists: Adapting the eagle-i Platform and a Proof-of-Concept Educational Resource Ontology

Juliane Schneider\textsuperscript{1}; Marc Ciriello\textsuperscript{1}; Jessica Zhang\textsuperscript{1}; Ankit Panchamia\textsuperscript{1}; Xavier Haurie\textsuperscript{1}; Thomas Naughton\textsuperscript{1}; Diane Keogh\textsuperscript{1}; Doug MacFadden, MS\textsuperscript{1}\
\textsuperscript{1}Harvard Medical School, Boston MA

Introduction
The Postgraduate Education Program in Clinical and Translational Science (PGE) at Harvard Medical School offers educational content through their Advanced Curriculum Compendium (ACC) and Educational Video Library (EVL). To improve discoverability of their resources, PGE sought to combine the two separate collections while making the resource data ontology-compliant for discovery via a novel search interface.

Making educational resources discoverable is an area of some effort in the Clinical and Translational science community, but there is no broadly-accepted solution, let alone a solution that follows semantic principles to enable data interoperability. The Harvard Catalyst team used the eagle-i platform adapted to use the N-Lighten educational resources ontology. The N-Lighten ontology was developed by Oregon Health & Sciences University in conjunction with Ohio State University to explore the possibility of developing an ontology for this domain area. The Harvard Catalyst team collaborated with PGE to 1) create the new search interface and 2) apply the N-Lighten proof-of-concept ontology in its first known practical application.

Methodology
Development and data requirements were documented with PGE after an analysis of the legacy systems and data. Desired search functionalities and filters for the user interface were identified. To leverage the N-Lighten ontology, both mapping of data from the legacy systems and selected ontology modifications were required. The N-Lighten ontology was extended to meet local needs by adding branches for the 0-5 Translational Levels, for course status (active/inactive), and for learning type (in-person or online). Ontological relationships (Part of/Has Part properties) were used to link videos to associated courses, thus connecting resources from the legacy systems.

Results
The N-Lighten proof-of-concept discovery tool was amended and the new search interface built according to local requirements. The new website-based search function as expected and a screenshot of the search interface is shown in Figure 1. Further development of the N-lighten ontology, if undertaken, may integrate the changes made to suit this local application to ease application to other institutions and uses.

Figure 1: Search interface with search box, filters, and sort features
**Simpler is Better: Case Studies of Clinical Decision Support for Laboratory Utilization/Stewardship**

**Thomas Schneider, MD**, **Ila Singh, MD PhD**1, **Bruce Darrow, MD PhD**1, **Joseph Kannry, MD**1

1Mount Sinai, New York, NY; 2Texas Children’s Hospital, Houston, TX

**Introduction:** Clinical Decision Support is best defined as anything that supports a decision.1 With that in mind CDS can involve other interventions in the EHR as simple as renaming order or changing the workflow. Clinical Decision Support (CDS) systems play a key role in Utilization Management/Stewardship programs.2 From 2014-15 an operational governance committee, the Laboratory Utilization Formulary Subcommittee, was charged with identifying tests that were clinically overutilized. The committee consisted of domain experts and EHR technical teams and was co-chaired by Informatics and the laboratory director. Seven laboratory tests were identified based on feasibility and suitability for EHR changes. The interventions were EHR test name changes, checking for duplicate test results in EHR, removal of tests from common order sets (collections of orders), removal from the list of orderable tests (aka Inpatient Preference List), and personal communications and information sessions with the biggest test utilizers. These interventions were applied to the following tests: Hemoglobin A1C, Hepatitis A antibodies, lipid panel, vitamin D, reticulocyte count, and a serum Iron + TIBC panel. Test utilization in the year prior to the intervention (2013) was compared to the year after the interventions with follow up evaluations a few years post-intervention in 2018. The interventions were restricted to inpatient laboratory test orders.

**Results:** For hemoglobin A1C, there were two interventions, a 90-day duplicate check and the removal of hemoglobin A1C from the insulin ordering pathway available to all inpatient clinicians. Vitamin D testing consisted of only two interventions, duplicate checking of less than 60 days and a name change. Vitamin D 1,25 Dihydroxy was changed to ‘Vitamin D 1,25 Dihydroxy (NOT for deficiency screening)’ and Vitamin D 25 Hydroxy was changed to ‘Vitamin D 25-OH (for deficiency screening).’ Hepatitis A antibody tests were renamed with “Hepatitis A Virus Antibody IgM” being renamed to “Hepatitis A IgM (Acute Infection)” and “Hepatitis A Virus Antibodies Total” being renamed to “Hepatitis A Virus Ab Total (immunization status).” Then more recently (2016) Hepatitis A Total Antibody was renamed to “Hepatitis A Virus Total Antibodies” and “Hepatitis A Total Antibody” was renamed to “Hepatitis A Total Antibody.” The standard lipid panel was subjected to two interventions: duplicate checking and personal communication sessions. Reticulocytes intervention consisted of only duplicate checking within 3 days of the order being placed. The two test panel, serum Iron + TIBC, had a 60 day duplicate check and was also removed from a daily

<table>
<thead>
<tr>
<th>Lab Test</th>
<th>Pre-Intervention*</th>
<th>Immediate Post-Intervention*</th>
<th>Delayed Post-Intervention*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemoglobin A1C</td>
<td>313.2 [16748/53471] 13% duplicates**</td>
<td>191.9 [9362/48783] 8% duplicates**</td>
<td>206.6 [10041/48583] 8% duplicates**</td>
</tr>
<tr>
<td>Vitamin D 1,25 Dihydroxy</td>
<td>16.8 [800/47578] 10% duplicates**</td>
<td>9.6 [466/48685] 9% duplicates**</td>
<td>3.7 [181/48500] 10% duplicates**</td>
</tr>
<tr>
<td>Hepatitis A Antibody Total</td>
<td>25.5 [1218/47592]</td>
<td>18.5 [898/48664]</td>
<td>N/A ±</td>
</tr>
<tr>
<td>Hepatitis A IgM</td>
<td>25.5 [1217/47592]</td>
<td>23.9 [1162/48664]</td>
<td>31.3 [1326/48753] ±</td>
</tr>
<tr>
<td>Lipid Panel</td>
<td>N/A±</td>
<td>N/A±</td>
<td>10.2 [493/48500]</td>
</tr>
<tr>
<td>Reticulocytes</td>
<td>198.5 [9450/47611] 12% duplicates**</td>
<td>177.6 [8642/48664] 11% duplicates**</td>
<td>182.0 [8830/48504] 10% duplicates**</td>
</tr>
<tr>
<td>Serum Iron + TIBC</td>
<td>131.5 [6263/47611]</td>
<td>131.3 [6390/48664]</td>
<td>135.5 [6574/48504]</td>
</tr>
</tbody>
</table>

*Orders normalized to per 1000 discharges, in brackets are the number of discharges

**Conclusion:** Our case studies have demonstrated both successes and failures in clinical decision support in helping laboratory utilization. Laboratory test name changes were easier to implement compared to order set modifications. Duplicate checking led to variable successes, and clinician education led to a modest and sustained effect. Our largest success was in changing test names to those that added information needed to make the ordering decision.

**References**

Review of Medical Device Nomenclatures and Classification Systems

Martin Schotel,1 Karen M. Bavuso, RN, MSN,1 Melinda Wong, BS Pharm, RPh,1 Emily Chan, PharmD, RPh,1 Saverio Maviglia, MD, MS,1,2,3 Roberto A. Rocha, MD, PhD1,2,3

1Semedy, Inc., Waltham, MA; 2Harvard Medical School, Boston, MA; 3Brigham & Women’s Hospital, Boston, MA

Introduction

Different nomenclatures and classifications are used to categorize and identify medical devices. More than 20 nomenclatures and classification systems around the world are utilized to classify healthcare products and specifically, medical devices.1 Each nomenclature and classification system is typically utilized by different types of professionals based on specific needs, including procurement, adverse event reports, tariffs, among others.

An important difficulty resulting from these multiple nomenclatures and classifications is the inconsistency of the information that is ultimately captured, despite the growing need for more precise and reliable information about medical devices. This difficulty creates substantial overhead not only for healthcare institutions, but also for any medical device stakeholder, including providers, researchers, public health officials, and consumers (patients). The objective of identifying and tracking medical devices is essential when trying to determine which devices have been associated with adverse events, as well as efforts to coordinate and standardize care across institutions and settings.

Methods

We reviewed commonly used medical device nomenclatures and classifications, such as the Global Medical Devices Nomenclature System (GMDN),2 and the Universal Medical Devices Nomenclature System (UMDNS).3 Additionally, we reviewed a survey from the World Healthcare Organization (WHO) to identify device nomenclatures and their use.4 The intent was to identify properties used to describe, classify, and track devices.

Results

The WHO survey from 175 countries confirmed that 38 use GMDN or UMDNS, while 46 use a locally developed system. Six countries use multiple systems, while 85 do not use any classification.9 These variations were caused by multiple factors. For example, the FDA requires manufacturers to assign classification codes, while other countries adopted different methods. We are creating a property map showing inconsistencies across medical device systems.

Conclusion

Increasing numbers of medical devices are being used in healthcare environments, which creates a mandate for a unification effort. A fundamental goal of having a unified coding system is to allow proper descriptions and consistent sets of relevant details to standardize communication among all stakeholders. A set of essential properties describing medical devices and their clinical application is a significant requirement for such solution. A software platform able to detect and resolve inconsistencies at scale across nomenclatures and classifications is an important tool for cataloging medical devices and their use in any given type of institution or role.

References

2. Global Medical Devices Nomenclature System (GMDN). Available at: https://www.gmdnagency.org/
3. Universal Medical Devices Nomenclature System (UMDNS). Available at: https://www.ecri.org.uk/
4. WHO survey. Available at: https://www.who.int/medical_devices/priority/mde_nomenclature/en/index2.html

* http://gamapserver.who.int/gho/interactive_charts/health_technologies/nomenclature/atlas.html
**Associating Chief Complaints with Electronic Health Record Activity to Decrease Provider Administrative Burden**

John Schrom, MPH, Vijan Joshi, MD, Colleen Bouey, Allison Gilmore, PhD
One Medical, San Francisco, CA

**Introduction:** Physician burnout has become a widespread issue, with electronic health record (EHR) use known to be a major predictor of burnout[1]. The burden of EHRs has become so problematic that patients are now reluctant to engage in medical encounters due to the overuse of technology in the exam room[2]. The average physician spends almost six hours per day interacting with their EHR[3]. Machine learning and data mining can decrease this administrative burden through smart shortcuts, based on predicted user behavior.

**Methods:** Clinical encounters were gathered from a large, national primary care clinic system. The chief complaint was parsed into UMLS concepts using QuickUMLS[4], a clinical named entity recognition tool. These concepts were combined with patient demographics and clinical activities during the encounter: lab orders, prescriptions, diagnoses, and referrals. The apriori algorithm was used to discover association rules between these concepts. Performance was evaluated on a holdout dataset by removing a random concept and looking for a discovered association above a varying confidence threshold between the remaining concepts and the removed one. A control model was created by identifying the Top N most common diagnoses, prescriptions, and labs, and was similarly evaluated by looking for a removed concept in a Top N list, with varying N. The impact on physician administrative burden was estimated by number of clicks in the EHR, using the sample’s sensitivity and assuming one click reduction for medication and diagnosis, and two click reduction for lab orders.

**Results:** The dataset included 5,655,811 encounters, and association rule mining (ARM) discovered 3,019,661 rules. Many high-confidence rules involved direct correspondence of the chief complaint to the resultant activity (e.g., chief complaint of “bee sting” => new “bee sting” diagnosis). More nuanced associations — often symptom-based — also emerged (e.g., “broken condom” => STI testing; “motion sickness” => scopolamine prescription). Performance varied based on type of activity being predicted, with labs being the most performant (AUC = 0.69), followed by prescriptions (0.43) and diagnoses (0.41). The Top N models performed worse in all categories (see inset; labs: 0.36, prescriptions: 0.13, diagnoses: 0.14). We estimate this approach results in 3.2 fewer clicks per patient per year.

**Discussion:** This preliminary work demonstrates ARM as a method for identifying relationships between a patient’s demographics and chief complaint, and their provider’s EHR activity during the encounter. This approach performed better on lab ordering than medication prescribing or diagnosing. This may be due to larger number of classes being predicted, as well as more clinical variability in prescribing behavior and inconsistent diagnosis documentation. Even with its current performance, ARM is more predictive than a “Top 10 Most Common” style approach for the same activity. This suggests that implementing rules-based shortcuts in the EHR to allow single-click ordering, prescribing, and documenting would be more effective than listing the most common activity. Our approach would introduce opportunities to decrease clinical variation and improve documentation, while also saving the administrative burden of an estimated 4,800 clicks for a provider with a panel of 1,500 patients.

**References**
Neural Relation Extraction from Biomedical Literature
Elaheh Shafieibavani, PhD1, Antonio Jimeno Yepes, PhD1
1 IBM Research, Southbank, VIC, Australia

Introduction Genetic mutations are linked to diseases such as cancer and might be related to drug resistance. Curation from the literature into databases is typically done by human curators, which is slow and expensive. Text mining methods exist to identify mentions of mutation in text, such as T790M, which specifies a protein mutation in position 790. Position in a mutation mention is meaningless without a reference to a gene or protein. We propose linking mutations to the gene or protein mentioned in the scientific articles since this will provide a more complete contextualization of the mutations. In this case, T790M is linked to EGFR protein in MEDLINE document PMID:29448920.

Methods and Results We used the AMIA dataset available in1, with additional examples in the training set generated manually. The training/testing sets contain 2656/385 mentions of mutations and 2799/280 of genes/proteins and 1617/130 relations between genes and mutations. We extracted about %30 of our training set as the validation set. For further analysis and experiments, we also used the entity relations and chemical-protein datasets originated from the BioNLP2 and BioCreative3 VI shared tasks. Having mentions of genes/proteins/chemicals and mutations/diseases, the purpose is to automatically extract their relation within a sentence span. To this end, we have considered the following two approaches4: (i) Support Vector Machine (SVM) and (ii) Convolutional Neural Network (CNN), that model the input sentence as a linear sequence of word tokens. Each word token is then mapped to relevant vector space embeddings. The word vector, part-of-speech and entities features are directly produced from the information of each token. The distance and relative position features model the position of the token in the sentence. The path features mark the dependencies connecting each token to a token of interest. The shortest path features mark the set of dependencies forming the shortest path for a candidate relation. These embeddings are then concatenated together to be processed by a set of 1D convolutions with window sizes 1, 3, 5 and 7, and an applied global max pooling to each convolutional layer. The convolution output vector is fed into a dense layer of 400 neurons, which is connected to the final classification layer that uses sigmoid activation while the other layers use ReLU activation. Training multiple models with randomized initializations, we use the average prediction of the five best models ranked with micro-averaged F-score on the validation set. These ensemble predictions are calculated for each label as the average of all the models’ predicted confidence scores. In addition, a random combination of hyperparameters is picked from the ranges to be optimized. The adam optimizer with binary crossentropy is used with a learning rate of 0.001, dropout of 0.1, and up to 500 epochs, with early stopping once validation loss has no longer decreased for 10 consecutive epochs. Table 1 shows the performance of these approaches comparing against the human agreement for the AMIA dataset, and the corpora for the BioNLP and BioCreative relation extraction tasks.

Conclusion Experiment results show that it is possible to automatically extract the relations between mutations and genes/proteins with higher performance to human annotators from biomedical literature. Besides, we observe that the CNN F-score is superior to the SVM’s for different relation extraction tasks (AMIA, REL11, and CP17).

References


Predicting Acute Care Utilization (ER Visit and Hospitalization) due to Social Determinants of Health Factors

Soham Shah, MD MBA1,2,3, Peter Nguyen, MD1,2, Manisha K Shah, PhD3, Hamed Abbaszadegan, MD MBA FACP FAMIA1,2

1University of Arizona College of Medicine, Phoenix, AZ; 2Phoenix VA Health Care System, Phoenix, AZ; 3Mayo Clinic, Phoenix, AZ

Introduction: United States national health care expenditure was $3.5 trillion, representing 17.9% of the GDP [1]. There has been an increasing recognition of the influence of social, behavioral, economic, and environmental factors on overall patient health. It is suggested that social circumstances, behavioral patterns, and environmental exposures may have a combined contribution as high as 60% on an individual’s risk of premature death. Emergency room (ER) visits and hospital readmissions are part of indirect measures of health quality and have been a major focus due to their large contribution to healthcare costs. Studies have shown correlation between social factors and ER visits and 30-day hospital readmission [2,3]. Since social determinants of health disproportionately affect the population in lower socioeconomic status, strategies to improve healthy quality should focus on this population [4]. The rationale for this project then is to develop a more data-driven approach to proactively help patients with risk factors based on the social determinant of health list obtained from Arizona Health Care Cost Containment System (AHCCCS) [5]. The goal is to predict groups of patients at high risk from specific social factors for whom proactive care coordination and access to resources may be valuable.

Methods: The active patient population at the Phoenix VA Medical Center 10/1/2015 to 12/31/2018 were identified and a retrospective analysis was performed looking for ICD10 codes corresponding to social determinants of health risk factors. Annual number of ER visits and hospitalizations per patients were reviewed as outcome metrics. The social factors were based on a list published by AHCCCS [5]. Logistic regression models were used to identify correlations between the social determinants (ICD10 codes) and hospitalization or ER visits.

Results: Data from a cohort of 23,966 patients within the Phoenix VA system from different ethnicities, employment, and marital statuses were analyzed. The first outcome allowed for geographical analysis, identifying zip codes in the Phoenix Metro with increased acute care utilization. A standard correlation analysis between ICD10 codes and Hospitalization identified Z659 (problems related to psychosocial circumstances) and age as the top 2 and fulltime employment or active military as the least correlated events for hospitalization. Similarly, Z659 and Z596 (low income) as the top two and Marriage and Z642 (release from prison) as the lowest correlated for ER visits. Individually correlated ICD10 codes only yielded a maximum correlation coefficient of ~0.2. However, using machine learning and a logistic regression with ICD10 codes and demographics as inputs, our model was able to correctly predict no hospitalization 95.8% of time, and ER visits 78.8% of time.

Discussion: This work was able to predict no hospitalization 95.8% of time, ER visit 78.8% of time and correlate social determinants with acute care utilization, either through visits to the ER or hospitalization. The ability to predict these using machine learning would guide interventions towards a large patient population to prevent ER visits or hospital admissions. We believe this model can be used in real-time to facilitate better health outcomes and resource utilization, thus reducing overall cost of care while maintaining quality and value.

References
Enhancing Allergy Documentation in a Commercial EHR System
Sonam Shah1,2, Carlos Ortega1, Suzanne V. Blackley3, Kimberly G. Blumenthal4,5, Foster R. Goss6,7, Paige G. Wickner4,5, Diane L. Seger1,3, David W. Bates1,4, and Li Zhou1,4
1Brigham and Women’s Hospital, 2MCPHS University, 3Partners Healthcare, 4Harvard Medical School, 5Massachusetts General Hospital, 6University of Colorado Hospital, 7University of Colorado School of Medicine

Background: More than 20% of patients report at least one drug allergy. Adverse drug reactions (ADRs) occur in 1 in 4 outpatients and 1 in 5 inpatients. Complete and reliable patient allergy histories are critical to inform future prescribing. However, allergy documentation is often incomplete with reactions and reaction type specified only half of the time. Distinguishing reaction types (e.g., allergy vs intolerance) is important for patient care but challenging for general clinicians to determine. Another documentation obstacle is the reliance on commercial or local dictionaries whose adverse reaction lists are incomplete, ambiguous, static, and vary by location. Reactions lists from 6 institutions ranged from less than 20 to more than 100. EHRs support entry of structured/coded reactions, but free-text reactions remain common (1/6 of entries). We sought to improve reaction documentation beginning with the development of a more comprehensive, specific, and correctly classified reaction list.

Methods: We assigned a reaction type (i.e., allergy, contraindication, or intolerance) and minimum criticality (i.e., low, medium, or high) to each reaction in our institution’s reaction pick list. Allergy was defined as an immunologic hypersensitivity reaction of any possible Gell and Coombs type (i.e., Type I, Type II, Type III, Type IV). Intolerances were defined as development of non-immunologic symptom or sign from a substance. Contraindication was defined by a medical condition where a medication, food or other substance should never be used because of known harm that would occur. Reaction types were assigned by a clinical expert group consisting of pharmacists and physicians, including allergy specialists. Next, we used MTERMS, a multipurpose NLP tool for processing clinical text, to determine the prevalence of reactions including those entered as free-text. MTERMS extracted reactions from free-text comments using our previously developed value set of 4,300 reaction terms. The terms were then mapped to SNOMED-CT concepts, and the mappings were validated by clinical experts. The concepts were then further mapped to the EHR’s reaction list via exact match. Based on the frequency of free-text reactions without a corresponding coded reaction, clinical experts considered which reactions warranted coded entry in the existing reaction pick-list and classified reactions by type and criticality.

Results: The Partners reaction pick-list currently has 46 reactions (excluding “Other” and “Unknown”), with 26 of these reactions (56.52%) classified as allergies, 10 (21.74%) as intolerances, and 10 (21.74%) as contraindications. 24 reactions (52.27%) were considered high criticality, 11 (45.83%) were considered medium, and 11 (45.83%) were considered low. We identified 51 additional reactions with high frequency and clinical importance that warranted adding to the existing pick-list (Table 1). Of these, 22 (43.14%) were intolerances, 20 (39.22%) were contraindications, and 9 (17.65%) were allergies. Overall 34 of these reactions (66.67%) were high criticality, 8 (15.69%) were medium, and 9 (17.65%) were low.

Conclusions: Allergy documentation varies institutionally. We created an enhanced allergy reaction list informed by free text entries. We assigned reaction types and minimum criticality for all reactions to assist in improved allergy documentation. Future steps will include the implementation and evaluation of this enhanced reaction list.

Acknowledgement: This study is funded by Agency for Healthcare Research and Quality (AHRQ) R01HS25375R01

References:
National Working Group to Standardize the Identification of Sensitive Data Elements to Support Patient Privacy

Marianne Sharko, MD, MS1, Hannah Galvin, MD, FAAP2, Susan J. Kressly, MD, FAAP3, Joseph H Schneider, MD4, Fabienne Bourgeois MD, MPH5, Feliciano (Pele) Yu, MD6, Matthew Hong, MHCI7, Lauren Wilcox, PhD7, Jessica S Ancker, MPH, PhD1

1. Department of Healthcare Policy & Research, Weill Cornell Medical College, New York, NY, 2. Lahey Health, Burlington, MA, 3. Kressly Pediatrics, Warington, PA, 4. UT Southwestern Medical Center, Dallas, TX, 5. Harvard Medical School, Boston Children’s Hospital, Boston, MA, 6. Arkansas Children's Hospital, Little Rock, AR, 7. School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA

Problem addressed:
In this age of electronic medical information collection, storage, analysis and exchange, it is important to protect patients’ rights to privacy over sensitive medical information. State laws and medical center policies governing privacy protection can vary and provide a complex backdrop for the standardization of sensitive data. It is important to consistently identify sensitive information in order to prevent the unintentional sharing of information through the electronic health record (EHR), patient portal or health information exchange (HIE).

Without standardization, medical organizations are left on their own to define the types of information identified as sensitive, and in need of heightened protection. This results in varying lists of sensitive data which can leave patients vulnerable to breaches of privacy. This is of particular importance to vulnerable populations, such as adolescents, patients seeking mental health or sexual reproductive care, or patients using proxies to manage their healthcare. HIE can provide challenges to data protection due to differing privacy policies. With limited ability to identify sensitive data in the EHR, many systems rely on individual professionals to provide this protection. The informatics community and industry have found it difficult to address these challenges due to its complexity and lack of business and regulatory drivers. However, it continues to be a clinical problem that impacts patients and providers daily.

Purposes of the working group:
A previous study found that many chief medical information officers (CMIO’s) welcome more formal assistance in the identification and control of sensitive data; however, national guidelines should be thoroughly vetted. The established standards should align with recommendations from medical associations and state laws, and should be adaptive enough to accommodate differing patient needs. We have developed a working group through a snowball recruitment process to define the problem and to brainstorm potential solutions. This group, working with the AAP Council of Clinical Information Technology and the Child Health Informatics Center includes pediatricians, adolescent specialists, CMIO’s, informaticists, researchers, and human-computer interaction specialists.

A primary aim of the group is to develop industry and expert consensus around recommendations for privacy protections at the data element level, a granularity previously thought to be too complex to address, yet which is necessary to support clinical use cases. Currently a clear list of sensitive data does not exist. To that end, we have begun categorizing data elements into levels of sensitivity through a small group consensus process to identify those with heightened privacy protection needs, such as from mental or sexual histories. We have also developed clinical use cases to demonstrate the need for standards development. Communication with vendors to delineate privacy expectations and create functional standards will help motivate electronic advancements, provide long-term consistency, and garner support from the ONC through their rule-making process. We look forward to sharing our processes, sensitive data lists and use cases to advance this conversation so that patients and families can benefit from the accurate exchange of information with confidence that their wishes regarding data privacy are respected.

References:
Identification of Latent Subtypes of Patients with Opioid Misuse

Brihat Sharma, BA1; Dmitriy Dligach, PhD1; Robert Kania, BS1; Elizabeth Salisbury-Afshar, MD, MPH2; Niranjan Karnik, MD, PhD2; Cara Joyce, PhD1; Majid Afshar, MD1
1Loyola University Chicago, Chicago, IL; 2Rush University, Chicago, IL

Introduction: The aim of this study is to identify subtypes of opioid misuse using latent class analysis (LCA) on readily available structured electronic health record (EHR) data in hospitalized patients. Additionally, the clinical notes frequently contain free text about social and behavioral determinants of health that cannot be comprehensively examined manually due to high volume of data; therefore, topic modelling was applied to summarize the corpus of text. The goal is to identify distinct subtypes of patients with opioid misuse, and characterize the subtypes using topic modelling. We hypothesize LCA will identify distinct subtypes in our patient cohort with clinically meaningful traits.

Methods: This study utilized data from the EHR of an urban tertiary academic center between January 1, 2007 and September 30, 2017. An operational definition for opioid misuse was developed following the National Survey on Drug Use and Health criteria for opioid misuse. Among those with opioid misuse, latent classes were identified using demographics, diagnosis codes, and urine drug screen results, and model fit statistics were utilized to identify the appropriate number of latent classes. Pre-processing of all clinical notes was performed in the Apache clinical Text Analysis Knowledge Extraction System (cTAKES) with concept mapping to concept unique identifiers (CUIs). Latent Dirichlet Allocation (LDA) was used to identify the topics. The LDA model was trained on the full cohort of patients with opioid misuse and produced topics expressed as probability distributions over CUIs. The optimal number of topics was identified by examining a range of topics using model fit statistics for topic coherence.

Results: The health system had 228,884 inpatient encounters during the study period and 6,224 (2.7%) met inclusion criteria for opioid misuse. Fit statistics and model interpretability suggest a 4-class model was optimal. Class 1 represented high hospital utilizers with known opioid-related conditions (36.5%); Class 2 contained illicit users with low socioeconomic status and psychoses (12.8%); Class 3 consisted of those with alcohol use disorders with complications (39.2%); and class 4 had low hospital utilizers with incidental opioid misuse (11.5%). Classes 1 and 3 had highest likelihood for 30-day unplanned readmissions. In topic modelling, the final data corpus in the 6,224 patient encounters was comprised of 25,801 unique CUIs across 422,147 clinical notes. Twenty topics were identified for the cohort, with themes spanning from chronic medical conditions to behavioral conditions and healthcare services. LDA topics complemented the LCA findings (Figure).

Figure. Distribution of probabilities for each topic across the 4-class model

Conclusion: We identified a four-class model that identified clinically interpretable and relevant subtypes for opioid misuse, with good class separation and face validity based on documentation in the notes, structured data, and clinical outcomes. LCA using readily available class-defining substance use variables from the EHR may be applied as a prognostic enrichment strategy for targeted interventions (Research supported by NIH R01LM012973 Dligach).

References:
Do Urban and Rural Patients Use Telemedicine Differently?

Lincoln R. Sheets, MD, PhD¹, Emmanuelle Wallach², Mirna Becevic, PhD¹
¹School of Medicine, University of Missouri, Columbia, Missouri, USA;
²PhD student, College of Engineering, University of Missouri, Columbia, Missouri, USA

Introduction

Telemedicine has traditionally been used in rural areas to improve access to specialty care.¹ However, telemedicine use is growing rapidly in urban areas,² and little is known about differences in urban and rural telemedicine use. We studied whether rural and urban patients in Missouri use telemedicine to access healthcare specialties at different rates.

Materials and Methods

We obtained Institutional Review Board approval to study de-identified electronic health records (EHR) from telemedicine visits at the University of Missouri from 2008 through 2017. We used Federal Office of Rural Health Policy (FORHP) classifications of patient zip code as rural or urban.³ Visits were categorized according to provider specialty: dermatology, child psychiatry, adult psychiatry, and other specialties which accumulated to less than 5% of all visits (anesthesiology, child development, endocrinology, family practice, neurology, oncology, vascular surgery).

Results

Of 5,411 telemedicine visits, 34% were from urban areas; these were 73% child psychiatry, 10% adult psychiatry, 13% dermatology, and 4% other, while rural visits were 35% child psychiatry, 19% adult psychiatry, 44% dermatology, and 2% other. Figure 1 summarizes these findings and Table 1 shows counts and demographic characteristics of telemedicine visits in each specialty, some of which were statistically (p<0.05) different between rural and urban visits.

Table 1. Counts and demographics of rural and urban visits.

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Location</th>
<th>Total visits</th>
<th>Unique patients</th>
<th>Mean age at visit</th>
<th>Mean visits/patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child Psych</td>
<td>Rural</td>
<td>1,254</td>
<td>302</td>
<td>12.7*</td>
<td>7.07*</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>1,339</td>
<td>453</td>
<td>15.6*</td>
<td>2.91*</td>
</tr>
<tr>
<td>Adult Psych</td>
<td>Rural</td>
<td>665</td>
<td>178</td>
<td>42.3</td>
<td>3.83*</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>178</td>
<td>128</td>
<td>43.3</td>
<td>1.55*</td>
</tr>
<tr>
<td>Derm</td>
<td>Rural</td>
<td>1,582</td>
<td>900</td>
<td>39.2*</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>242</td>
<td>156</td>
<td>25.1*</td>
<td>1.55</td>
</tr>
<tr>
<td>All Other</td>
<td>Rural</td>
<td>81</td>
<td>36</td>
<td>39.6</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>70</td>
<td>39</td>
<td>39</td>
<td>1.79</td>
</tr>
</tbody>
</table>

* Significant (p<0.05) difference between urban and rural results.

Conclusion

Urban patients access telemedicine significantly differently, with a much greater demand for child psychiatry. Telemedicine may offer help for the critical nationwide shortage of adolescent psychiatry providers in rural and urban areas. Health systems should consider adjusting resources to meet the different needs of these two populations.

References

Leveraging a Data-Driven Strategy to Enrich Rare Disease Knowledge

Feichen Shen, Ph.D., Yiqing Zhao, Ph.D., Hongfang Liu, Ph.D.
Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Abstract

Existing resources to assist the diagnosis of rare diseases are usually curated from the literature that can be limited for clinical use. It often takes substantial effort before the suspicion of a rare disease is even raised to utilize those resources. The primary goal of this study is to apply a data-driven approach to enrich existing rare disease resources by mining phenotype-disease associations from electronic medical record (EMR).

Background

In this study, we focused on the Human Phenotype Ontology (HPO) [1] and Orphanet [2]. An annotation file that records associations between phenotypic terms from the HPO and rare disease terms from Orphanet were referred here as rare disease knowledge resource (HPO-Orphanet). We propose to enrich the HPO-Orphanet through mining association information between clinical phenotypes and diseases using EMR. Such enriched information, named as HPO-Orphanet+, can be used to link similar rare/common diseases and provide differential diagnostic decision support at the point of care for rare disease diagnosis.

Methods

We formulated the task of mining association information between clinical phenotypes and diseases as an association rule mining task. Specifically, patients are considered as transactions and their phenotypes and diseases are considered as items. The phenotype-disease association discovery can be defined as one item rule {Phenotype} → {Disease}. We first leveraged the support and confidence metrics to measure the importance of discovered associations. We then built a bipartite graph (HPO-Orphanet+ Graph) as an enriched knowledge graph to represent associations between diseases and phenotypes. We finally used Jaccard similarity on disease level to measure the similarity among diseases based on significant phenotypic features selected by odds ratio.

Results

All clinical notes during the years of 2010 to 2015 from Mayo Clinic EMR were used for the study. For each note type, we focused on the diagnosis section of the notes which summarizes problems for each patient. The resulting corpus contains 12.8 million clinical notes corresponding to about 730K patients. We first carried a use case study on Hodgkin lymphoma to compare performance on rare disease differential diagnosis for three different bipartite graphs, which are the original knowledge graph HPO-Orphanet, data-driven EMR Graph, and the enriched knowledge graph HPO-Orphanet+. We used ranked disease-disease similarity generated by the eRAM [3], an existing rare disease encyclopedia, as a gold standard to compare the three graphs with sensitivity and specificity as (0.17, 0.36, 0.46) and (0.52, 0.47, 0.51). We found HPO-Orphanet+ yielded the highest sensitivity. We also compared the top 15 diseases generated by HPO-Orphanet+ with eRAM and another widely used clinical diagnostic tool, the Phenomizer [4]. We found that the HPO-Orphanet+ graph identified 46.7% (7 out of 15) common diseases and 53.3% (8 out of 15) rare diseases. While differential diagnostic candidates provided by the Phenomizer are all rare disease. Similarly, the eRAM generates 93.3% (14 out of 15) rare diseases but only 6.7% (1 out of 15) common diseases. Compared to the Phenomizer and eRAM, the HPO-Orphanet+ graph is more capable of detecting link between common and rare diseases at the early time of diagnosis to assist in diagnostic decision support.

Conclusion

In this study, we proposed a data-driven approach to mine phenotype-disease associations buried in EMR so as to enrich current rare disease knowledge with extracted associations as well as differential diagnostic suggestions.

References

Extracting Disease Onset from Family History Comments in the Electronic Health Record using Fast Healthcare Interoperability Resources

Jianlin Shi MD, PhD\(^1\), Kensaku Kawamoto, MD, PhD, MHS\(^1\), Wendy Kohlmann, MS\(^2,3\), Danielle L. Mowery, PhD\(^4,5\), Richard L. Bradshaw, PhD, MS\(^1\), Subhadeep Nag, MS\(^1\), Wendy W. Chapman, PhD\(^1\), Guilherme Del Fiol, MD, PhD\(^1\)

\(^1\)Biomedical Informatics, \(^2\)Population Health Sciences, University of Utah, UT
\(^3\)Huntsman Cancer Institute, Salt Lake City, UT
\(^4\)Biostatistics, Epidemiology, & Informatics, \(^5\)Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA

Abstract Familial risk criteria for common cancers rely on the age of cancer onset, which is documented inconsistently using structured and unstructured formats in electronic health records (EHRs). We operationalized a natural language processing (NLP) system to extract disease onset information from free-text EHR fields with recall and precision ranging from 98% to 100%.

Introduction

Family health history (FHH) is useful for identifying patients with increased cancer risk. Evidence-based criteria for patients who may benefit from the genetic evaluation include the age of cancer onset in family members (e.g., first or second degree relative with breast cancer and age of onset (AOO) <45 years old). However, the AOO is often recorded in free-text and is not readily available for standardized health information exchange and computational purposes. We investigated a natural language processing (NLP) approach to (i) extract AOO from a free-text comment field in an electronic health record (EHR); and (ii) structure the AOO to comply with the Fast Healthcare Interoperability Resources (FHIR) standard to allow inter-organization information exchange.

Methods

We queried the University of Utah Health (UHealth) Enterprise Data Warehouse for all patients aged 25-50 seen at least once at UHealth primary care clinics between April 2014 and May 2018. We selected each patient’s most current FHH entries within the EHR containing documentation of breast, colorectal, endometrial, and unspecified cancer. Each FHH entry has structured data on the diagnoses, relation, and AOO (if completed) with more detailed information in the comments field. We applied a rule-based AOO identifier built on EasyCIE\(^1\) (a user-friendly clinical NLP tool) to extract AOO from the comments. Next, we instantiated the extracted AOO using the FHIR FamilyMemberHistory resource, which allows representing AOO in one of four data types: numeric, range, period, or string. Two annotators manually reviewed a random sample of 300 positive (i.e., with age of onset) and 300 negative (i.e., without age of onset) cases. We then calculated sensitivity (recall), specificity, and positive predictive value (precision) for each FHIR data type.

Results

The sensitivity, specificity, and PPV achieve between 98% to 100% across all four data types. Of the non-flagged cases, we observed only one range error.

Table 1: Disease age of onset classification by data type. PPV = positive predictive value.

<table>
<thead>
<tr>
<th>Age of Onset Data Types</th>
<th>Reference Standard</th>
<th>EasyCIE</th>
<th>Sensitivity</th>
<th>PPV</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>143 (48%)</td>
<td>140 (47%)</td>
<td>100%</td>
<td>98%</td>
<td>98%</td>
</tr>
<tr>
<td>Range</td>
<td>120 (40%)</td>
<td>123 (41%)</td>
<td>98%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Period</td>
<td>36 (12%)</td>
<td>36 (12%)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>String</td>
<td>1 (&lt;1%)</td>
<td>1 (&lt;1%)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Discussion

The most spurious false-positive errors for age were caused by a missing modifier (“?”” indicating fuzzy range). The only false-negative error was a missed range: “70s I think”. We are enhancing EasyCIE to extract and encode additional FHH information, including family relation, sidedness, and cancer type using FHIR.

Acknowledgments

This research was supported by grant number U24CA204800 under the Informatics Technology for Cancer Research (ITCR) program of the National Cancer Institute of the US National Institutes of Health.

\(^1\) https://github.com/jianlins/EasyCIE gui
The impact of emotion on medication errors: A text mining analysis approach

Akiko Shibuya, RN, PHN, PhD1, Jimpei Misawa, PhD2, Rie Ichikawa, MD, PhD2, Yukihiro Maeda, BH1, Izumi Ikeda, RN, PHN, MSN1, Mihoko Koshida, RN, PHN, PhD1, Teruyoshi Hishiki, MD, PhD3, Yoshiaki Kondo, MD, PhD2

1 Faculty of Nursing, Toyama Prefectural University, Japan; 2Department of Health Care Services Management, Nihon University School of Medicine, Tokyo, Japan; 3Department of Information Sciences, Toho University, Chiba, Japan

Abstract
The initial phase of decision-making before any medical procedure is largely affected by the emotions of healthcare providers. We used a text mining approach to analyze the emotions of healthcare providers during decision-making prior to drug treatment to determine the association between emotions and medication errors. Our approach may reveal how the emotions of physicians affect medication errors, and may have a significant impact on patient safety.

Introduction
Medication errors that occur in hospitals have been a source of concern for decades. There is an increasing number of studies focusing on the analysis of medication errors. Studies suggest that errors in medication dose and usage are common, and that these errors result from the lack of attention, verification, and knowledge of healthcare providers1. However, only a few studies to date have focused on the emotions of healthcare providers and how they affect decision-making prior to any medical procedure. Some studies examined human factors in medical errors2; however, these studies focused on the behavior and experience of healthcare providers, and only a limited number of studies focused on emotions and used text mining approaches to analyze medical errors. In the present study, we focused on the emotions of physicians during the decision-making phase prior to drug treatment and used a text mining approach to analyze these emotions in the decision-making phase.

Methods and results
A total of 973 cases of medical errors were identified from the Japanese national medical error database using the terms “medication” and “physicians” for the search. Text mining was used to analyze emotions extracted from the comments sections that provided a summary for each medical error case. The association between medication error and emotion was subsequently examined. All analyses were performed using SPSS Text Analytics for Surveys 4.0 and SPSS 24.0. There were 764 negative comments and 228 positive comments (including some overlaps, p<0.001).

Figure 1. Emotions in medication errors

Figure 2. Association between medication errors and emotions

Negative emotions included fear, disgust, sadness, surprise, and anger (including some overlaps). Fear was the most common emotion (58.6%), followed by disgust (28.7%) and sadness (9.2%) (p<0.001). Overdosing was the most common error (15.6%), followed by wrong dose (13.1%) and wrong medication (7.4%) (p<0.001). Figure 1 shows the type of emotions for each error. A correspondence analysis was performed to determine the association between the medication errors and physicians' emotions (Figure 2).

Conclusion
We focused on physicians' emotions in the decision-making phase prior to any medical procedure, and used text mining to analyze physicians' emotions in the decision-making phase of drug treatment. The proposed method to determine the association between medication errors and emotions may help improve safety measures for drug treatment.

References
Predicting Patient Deterioration Using Continuous Monitoring and Concepts from the Field of Sports

Zvika Shinar, PhD¹, Veronica Maidel, PhD¹, Patricia C. Dykes, PhD, RN, FAAN, FACMI², Dalia Argaman, MSc¹, David W. Bates, MD, MSc²
¹EarlySense Ltd, Ramat Gan, Israel; ²Partners HealthCare System, Inc., MA, USA

Abstract
This study employed for the first time, continuous monitoring in non-ICU units, to predict major clinical deterioration. Applying machine learning on features extracted with the help of concepts from the field of professions sports, outperformed the commonly used MEWS score.

Introduction
Clinical deterioration of general ward patients can result in ICU transfer and/or death. Previous studies showed that continuous monitoring may provide a benefit in terms of improved patient outcomes and cost efficiency. The goal of this study was to test whether machine learning models based on features from continuous (contact-free) monitoring data and concepts from the field of sports¹ can improve standard Early Warning Scores (such as MEWS).

Methods
All patients hospitalized in five medical/surgical (non-ICU) units between April 2016 and April 2017 in a community hospital in Massachusetts, were monitored continuously with a contact free system (EarlySense Inc.). Monitoring included heart rate (HR), respiratory rate (RR), motion, and bed presence, every 10 seconds. Age-normalized features (based on maximal HR, calculated as 220-age, and HR range) were calculated and compared against a threshold corresponding to HR zones, as utilized in the field of professional sports (see Fig 1). Similar definitions were used for the first time for respiration. Machine learning classifiers (RFC & SVM) were trained to perform selection of the best features and to derive two predicting models, for expiration and for ICU transfer 24 hours prior to an event. The models were compared in a validation cohort using the area under the receiver operating characteristic curve (AUC).

Results
The study covered 11,737 unique hospital stays, including 138 deaths, and 770 ICU transfers. Validation model AUCs outperformed or were as good as MEWS AUC for the same data set. For ICU, the model’s AUC vs. MEWS’ AUC measures were 0.73 vs. 0.68 respectively. For mortality, MEWS and the model performed equally well, at 0.88.

Conclusion
We believe this work is novel in three ways: (1) we developed an early warning score based on continuous monitoring, (2) applied professional sports terminology to hospital medicine using an individual calibration of HR based on age and resting baseline, and (3) offer an analogy for the calibration of respiratory rate. The results outperform or perform as well as the commonly used MEWS score.

References
Multivariate Pathway Enrichment Analysis for Interpreting Expression Profiles

Ali Shojaee Bakhtiari 1, Kristin Wallace 2, Alexander V. Alekseyenko1, 3

1 Biomedical Informatics Center, Department for Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina, 29425, USA
2 Department of Public Health Sciences, College of Medicine, Medical University of South Carolina, Charleston, SC, 29425, United States
3 Department of Oral Health Sciences, Medical University of South Carolina, Charleston, South Carolina., 29425, USA

Abstract: Methods for extracting relevant biological information from expression profiles mostly rely on individual genetic expressions. Consequently, these methods are suboptimal in detecting small differences within the pathways that in aggregate result in a large multivariate effect. We propose Multivariate Pathway Enrichment Analysis (MPEA). MPEA harnesses the power of multivariate association to identify the subtle associations between genetic pathways and phenotype of choice. A reference is available at https://github.com/alekseyenko/MPEA.

Method: MPEA requires a multivariate expression array of control and test groups as well as reference to the genes involved in each genetic pathway. MPEA algorithm consists of three steps. Firstly, it calculates an enrichment score (ES) for individual pathways using distance correlation [2] of the expression profiles of the pathway between the control and test subjects. Next, it uses a permutation test over the entire expression profile to estimate the significance of the ES score. Finally, it corrects for multiple hypothesis testing to compensate for the multiple studies pathways.

Results: We compared the statistical power of MPEA against GSEA [3] for different effect sizes. Our simulation results show that MPEA has higher sensitivity compared to GSEA for different effect sizes, Figure 1. Next, we applied MPEA on real datasets and observed that MPEA is capable of identifying enrichment in pathways that are dimmed not significant in GSEA model. We bring an instance of the PCA plot of the expression profile of such pathway in Figure 2.

Conclusions: We showed that taking into consideration the interrelationship between the genes in genetic pathways significantly improves the statistical power of pathway enrichment detection.

Automated Information Extraction to Support Response Assessment in Myeloproliferative Neoplasms

Evan Sholle, MS1, Spencer Krichevsky, BS2, Sajjad Abedian, MS1, Prakash Adekkanattu, PhD1, Diana Jaber, BA2, Niamh Savage, BA2, Joseph Scandura, MD, PhD3, Thomas R. Campion Jr, PhD1,4,5

1Information Technologies and Services Department, Weill Cornell Medicine, New York, NY; 2Joint Clinical Trials Office, Weill Cornell Medicine, New York, NY; 3Department of Medicine, Weill Cornell Medicine, New York, NY; 4Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY; 5Department of Pediatrics, Weill Cornell Medicine, New York, NY

Introduction

The myeloproliferative neoplasms (MPNs) are a clinically related set of neoplastic malignancies of hematopoietic stem cells, pluripotent cells that give rise to multiple cell lineages in the blood. The primary criteria for effectiveness of MPN treatments is response, which is determined by a structured set of criteria issued by working bodies in the field. These criteria include, among others, myeloblast counts, transfusion dependence, cytogenetics, and mutational status.

The current gold standard for determining response in clinical trials requires manual review of the patient’s electronic health record (EHR) by trained personnel, a time-consuming and difficult task. While some of the data required for the assessment of response exist in a structured fashion, some exist in free text reports, necessitating review by trained staff to extract elements required to assess response. We hypothesized that by leveraging natural language processing (NLP) information extraction techniques, we could obviate the need for manual abstraction of these concepts.

Methods

Utilizing Leo, an open-source NLP platform developed at the Veteran’s Administration on the Apache Unstructured Information Management Architecture (UIMA), we developed an information extraction pipeline designed to extract three principal components from bone marrow biopsy reports: myeloblast counts, cellularity, and fibrosis. After initial pipeline development, trained research personnel annotated a sample of 300 bone marrow biopsies to serve as the gold standard. We then compared the output of the pipeline to manual annotation to assess its performance, calculating precision, recall, and F-score for each of the three components, as well as accuracy for extracted values. All calculations were performed using the R statistical software environment.

Results

The pipeline extracted aspirate myeloblasts with precision of 0.97, recall of 0.79, and F-score of 0.87; biopsy cellularity with 0.84 precision, 0.94 recall, and an F-score of 0.89; and fibrosis scores with precision of 0.91, recall of 0.91, and an F-score of 0.90. Extracted aspirate myeloblast scores were accurate in 95% of extracted cases. Extracted quantitative cellularity values were accurate in 73% of extracted values, and qualitative values were accurate in 56% of extracted values. Fibrosis scores were accurate in 69% of extracted values.

Discussion

While the assessment of response in cancer patients is a notoriously difficult problem, to our knowledge, this is the first documented effort to extract structured data elements from free text reports specifically for the purpose of supporting response assessment. This information extraction pipeline still holds the potential for improvement, especially with regard to extraction of qualitative assessments of cellularity. Future work in this area will revolve around incorporating the results of this pipeline into a larger technique for computational determination of response.

Acknowledgements

This study received support from NewYork-Presbyterian Hospital (NYPH) and Weill Cornell Medical College (WCMC), including the Clinical and Translational Science Center (CTSC) (UL1 TR000457) and Joint Clinical Trials Office (JCTO).
Information Seeking Practice Among Healthcare Professionals in Pakistan: A Review of Current Trends and Needs

Amna A. Siddiqui¹, Syed A. H. Rizvi¹, Jawwad Ahmed¹, Asif A. Syed², MD, MPH, Paul Fontelo³, MD, MPH. ¹Dow University of Health Sciences, Karachi, Pakistan; ²University of Illinois, Chicago, IL, USA; ³National Library of Medicine, Bethesda, MD, USA.

Introduction:
Evidence-Based Medicine (EBM) integrates clinical experience with current information to provide effective healthcare. We wanted to analyze the information requirements, current needs and prevailing trends of practicing doctors and medical students of a lower-middle-income country with only 33% of the population with internet access¹.

Methods:
Data was gathered from 160 participants (physicians and med-students) using convenience sampling by completion of either a paper or online questionnaire, which assessed their clinical background, technology access, need for health-related information, and the preference of resources to obtain information in different scenarios. Results were analyzed using SPSS version 17.0.

Results:
Basic mobile phones (with limited internet connectivity) is the most common device used at home (99.3%) and work (88.1%). Smartphones were available only to 17.5% participants at work. Internet connectivity was available for 96.8% at home but only 73.7% at work. Around 30.6% encountered questions in practice 2–4 times a day, and 50% were very likely to look up a reference. Most common resource for the majority of given clinical scenarios was a senior colleague, while medical websites (Medscape, Up-to-Date, WebMD etc.) were the first choice for non-specific general medical queries (Fig.1). Approximately 68.8% said that they used PubMed in daily practice. The most common reason for not using PubMed mentioned the ease in using other search engines, like Google.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug indications and dosage</td>
<td>Hard-copy of book, textbook, or journal</td>
<td>Medical website (Medscape, UpToDate, WebMD etc.)</td>
<td>Search Engine (Google, Yahoo, Wikipedia etc.)</td>
</tr>
<tr>
<td>Therapeutic management</td>
<td>Senior Colleague/Attending/Consultant</td>
<td>Medical website (Medscape, UpToDate, WebMD etc.)</td>
<td>Search Engine (Google, Yahoo, Wikipedia etc.)</td>
</tr>
<tr>
<td>Diagnostic tests and workup</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prognosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recommendation to patients</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General medical question</td>
<td>Medical website (Medscape, UpToDate, WebMD etc.)</td>
<td>PubMed / Medline</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Preferred source of information in given scenarios.

Conclusion:
Limitations of technological access must be considered when designing information resources in lower-middle income countries. As suggested by the participants, open and free internet connectivity, easy access to online resources, and tools with functionality independent of internet may enhance the uptake of EBM in Pakistan.

References

1762
Preventing inpatient hypoglycemia: A predictive model built from multi-center EHR data

Michael Simonov, MD1, Silvio E. Inzucchi, MD2, Francis P. Wilson, MD, MS1
1Yale Program of Applied Translational Research, New Haven, CT; 2Section of Endocrinology, Yale School of Medicine, New Haven, CT

Introduction
Severe hypoglycemia (blood sugar < 40mg/dL) is an important inpatient adverse event which is associated with increased morbidity, mortality as well as length of stay1,2. Centers for Medicare and Medicaid Services (CMS) is developing a hypoglycemia quality measure suggesting the importance of approaches to reduce such episodes. Current medical practice is heavily reactionary, namely a hypoglycemic episode prompts a subsequent change in medical care. There remains a need for proactive approaches aimed at predicting hypoglycemia before it occurs and prompting appropriate care. As an introductory step to changing such approaches, we set out to build an effective and implementable predictive model using existing health care record data with the ultimate goal of ready EHR implementation.

Methods
Inpatient data was obtained from three distinct hospitals in the Yale New Haven Health System (YNHHS). Data included time invariant data, such as demographics and past medical history, as well as time-updated data including vital signs and laboratory values. Severe hypoglycemia within the next 24 hours was set as the primary goal of the models. Several predictive models were evaluated; a primitive approach using only most recent glucose, a simple decision tree, and a multivariate logistic regression. Logistic regression and tree models were first trained on a subset of the data and then tested on the remaining held out data subset with performance measured by ROC curves.

Results
The final cohort included 203,962 patient encounters patient encounters with a total of 18.8 million individual time points for analysis. There were 2,633 encounters (1.3%) where the patient experienced at least one episode of severe hypoglycemia. Approximately two-thirds of these encounters were in diabetic patients and the remaining were in non-diabetics. Logistic regression was applied to data within the training dataset and the model displayed a performance of AUC 0.79 over the test data. In this logistic model, the most important features, as measured by z-value, were potassium, diabetes, minimum glucose in 72 hours, and bicarbonate. An easily-implementable decision tree model with four decision points displayed similar performance to the logistic regression model with decisions based on minimum 72-hour glucose, most recent glucose, creatinine, and length of stay. The primitive model, which was developed to align with current clinical practice, displayed very poor performance with an AUC of 0.55.

Conclusions
Logistic regression as well as a simple decision tree have good performance at predicting severe hypoglycemia. These models are superior to predicting hypoglycemia based solely on glucose, as is a common current clinical approach. We purposely limited model complexity by including readily-accessible and measurable features to construct parsimonious models which could be built in existing EHR decision support frameworks. The models presented here show promise in the development of a proactive approach for the prevention of severe inpatient hypoglycemia.

References
Adverse Events Monitoring for Medication Assisted Treatment of Opioid Use Disorder: Using Healthcare Data Interoperability to Inform Practice

Shyamashree Sinha, MD, MBA, MPH1, Robert Lee, MD1, Sarah Mullin, MS1, Jinwei Hu, MD1, Arlen Brickman, MD1, Angie Li, MD1, Peter L Elkin, MD, MACP, FACMI1
1Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, SUNY

Abstract

The goal of this study is to use electronic healthcare records from clinics and inpatient services to monitor opioid addicted patients who are on Medication Assisted Treatment (MAT), for adverse events including withdrawal, toxicity, and overdose. We also sought to find the treatment outcomes with methadone and buprenorphine in terms of the number of adverse events after the initiation of therapy.

Introduction


Retrospective patient records were collected from encounters occurring between January 2010 and April 2018 for patients with prescription opioid medications for non-cancer pain conditions after obtaining necessary IRB approval.

Method

The outpatient practice included 14,801 unique patients and the county hospital and outpatient clinic had an additional 8,330 unique patients. Electronic data from the different clinical data warehouse sources were collected, merged, de-identified, and then placed on a secure server. Deterministic matching was performed using social security number, date of birth, last name, and the first three letters of the first name. We were able to link 715 patients’ records between the two data sources. We examined the data for evidence of opioid related adverse events. These included withdrawal events, overdose, or other toxicity related events. Opioid related adverse events were defined by ICD-9-CM and ICD-10-CM codes for abuse, adverse events, complications, or poisoning. The outcomes examined did not include patient deaths. We then looked at the patients who had been prescribed an opioid medication, with or without a record of adverse events, who were subsequently placed on methadone or buprenorphine therapy based on National Drug Codes (NDC) of these medications.

Results: A total of 1096 patients on buprenorphine, 338 patients on methadone. 45.5% of patients were female, 87.2% were white, and 3.5% were Hispanic. 796 patients were removed due to a prescription date after diagnosis date. 229 of the buprenorphine group and 250 of the methadone group had an adverse event. The relative risk of having an adverse event while on methadone as compared to buprenorphine was 3.54 (95% CI 3.1040 to 4.0371) with a significance level of P < 0.0001. Number needed to treat (NNT) was 1.884.

Discussion: The findings of this study clearly indicates the advantage of buprenorphine or a combination therapy over methadone in having adverse events. Further longitudinal studies have to be performed to compare different aspects of MAT in the future.

Figure 1: The bar graph shows the number patients taking buprenorphine and methadone therapy. (Series1: Number of patients with adverse events before and after initiation of therapy; Series2: Number of patients with no adverse events reported.)


Acknowledgements: NIH NCATS Clinical and Translational Science Award (UL1TR001412-01); NIH NLM Training Grant T15LM012495-01 and NIH T322T32GM099607-06
Duplicate Laboratory Testing in the Context of Patient Care Transitions

Liron D. Sinvani, MD, Kayla Finuf, MS, Michael Qiu, MD, PhD,
Renee Pekmezaris, PhD, Jan Horsky, PhD
Center for Research Informatics and Innovation, Northwell Health, Manhasset, NY

Introduction

Timely and appropriate testing is critical for practicing evidence-based care and following clinical guidelines and best practices. The number of laboratory tests performed in the United States is estimated to be 4–5 billion annually. Over- and under-utilization of tests, however, can lead to lower care quality and higher costs. For example, unnecessary blood draws increase patient discomfort and the potential for infections and also the likelihood of false-positive results that lead to unwarranted interventions with further risks. Care transitions are known sources of patient risk from errors such as medication discrepancy and incomplete records but may also lead to laboratory repeats at intervals that are shorter than recommended. The objective of this descriptive study was to find the frequency of inappropriate repeat testing at a large, integrated healthcare delivery network for five commonly ordered laboratory tests. Specifically, we wanted to estimate the rate of duplicative testing in the context of care transitions between hospital and the community.

Methods

We queried a clinical data repository that aggregates all inpatient and ambulatory information from EHRs to find patients with at least one order during a 3-year period (2016-18) for the following common lab tests: TSH (thyroid stimulating hormone), HbA1c (Hemoglobin A1c), Cholesterol–lipid profile and for Vitamins B12 and D (25-Hydroxy). The collected dataset included order date and setting (hospital vs. community). Orders were marked as “repeat” when the time elapsed since the preceding test was shorter than guideline recommendation; otherwise as “index.” We analyzed testing rates and focused on the frequency of duplicate tests associated with care transitions, defined as scenarios where the index and repeated orders originated in different settings (hospital or community).

Results

There were 81,561 patients with 163,591 TSH orders, out of which 22,116 (14%) were repeats (<60 days), 81,205 patients with 179,706 HbA1c and 26,727 (15%) repeats (<90 days), 76,861 patients with 170,147 Cholesterol and 23,655 (14%) repeats (<90 days), 31,908 patients with 49,569 Vitamin B12 and 14,222 (29%) repeats (<365 days), and 44,629 patients with 90,735 Vitamin D and 33,137 (37%) repeats (<365 days). Repeat tests that followed a care transition totaled 3,168 (14%) for TSH, 6,122 (23%) for HbA1c, 3,993 (17%) for Cholesterol, 1,504 (11%) for Vitamin B12 and 2,488 (8%) for Vitamin D. The repeat rate was higher in the “to hospital” direction, from 1.4 times (B12) to 2.9 times (Cholesterol) than in the “to community” direction (Figure 1).

Figure 1 Tests repeated on care transition

A significant proportion of tests, ranging from 14% for TSH to 37% for Vitamin D, appeared to be duplicates. The high rates are congruent with previous published reports but we also found that duplicate tests were performed at up to 3-times higher rate (Cholesterol) during and after hospital admission than after discharge to the community. We hypothesize that when hospitalists have a problematic, laborious or nonexistent access to outpatient records they order the tests they need even if they may be repeats. EHRs where lab results are “silod” in displays and do not afford effective review of lab order history or when they are distributed over non-integrated systems likely contribute to test overutilization. Summaries of lab results medications that are available in discharge documents perhaps more conveniently aggregate this information for ambulatory clinicians who therefore have a better understanding of recent testing. Importantly, test duplication causes extra discomfort and pain to tens of thousands of patients every year and likely contributes to lower care satisfaction. Electronic alerts to clinicians about duplicate tests during ordering may help alleviate this problem. We plan to expand this study to include EHR interface evaluation and human-computer interaction studies to explore whether better system integration and usability can decrease duplicative testing.

The Cancer Imaging Archive (TCIA), NCI’s Imaging Data Resource

Kirk E. Smith, B.S.¹, Bill Bennett, B.A.¹, Lawrence Tarbox, Ph.D.¹, Tracy Nolan, B.E.¹, Geri Blake, M.Ed.¹, Justin Kirby, B.S.², John Freymann, B.S.², Fred W. Prior, Ph.D.¹
¹Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR
²Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Frederick, MD

Introduction
The Cancer Imaging Archive (TCIA) is the National Cancer Institute’s (NCI) repository of publicly accessible cancer images and supporting metadata¹. The data in TCIA are organized into collections by cancer type or research protocol. Collections are treated as data publications, with each one assigned a Digital Object Identifier (DOI) and citation to reference the dataset². Each collection includes a descriptive summary, licensing details, multiple search and download interfaces, citations, version control, and links to supporting data. TCIA also supports submission of third-party analyses of existing collections such as tumor segmentations and radiomic features. A help desk provides email and phone support along with user manuals and FAQs.

Data Acquisition and Curation
A TCIA submissions team consists of a Submissions Manager, a DICOM curator, and a Data Support Specialist who follow a standards-based approach to ensure accurate and timely acquisition and curation of data from submitting sites around the world. The Posda toolkit³ is used to inspect and edit each image for DICOM conformance, linkages to supporting data, PHI within the pixels, and PHI within DICOM elements. Care is taken in removing PHI to retain data that is useful to the researcher. A private tag dictionary is maintained within Posda. Once fully curated the data are moved into TCIA for public consumption.

Data Collections
TCIA supports 92 collections derived from major research initiatives including the Cancer Genome Atlas (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC), Quantitative Imaging Network (QIN), the Cancer Moonshot and clinical trials. TCIA also hosts the Lung Imaging Database Consortium (LIDC) collection, the Reference Image Database for Evaluation of Response (RIDER) collections and a number of NCI internal research programs as well as community initiated datasets.

Utilization Metrics
The success of TCIA may be measured by its number of users and published manuscripts. TCIA typically supports 15,000 active users per month from more than 125 countries that download approximately 75 TB of data. Over 700 publications have referenced TCIA and used its data. TCIA data been used in 14 image analysis challenge competitions⁴.

Conclusion
TCIA is a persistent archive that increases public availability of high quality cancer imaging data, supports NIH data sharing requirements, facilitates reproducibility of research, and creates a culture of open data sharing among collaborators. TCIA provides a number of services to support the submission and hosting of the data, relieving the contributors from the majority of data sharing risks.

References
Sentiment analysis methodologies of patient narratives: A descriptive study

Andrey Soares, PhD¹, Heather Coats, PhD, APRN-BC¹, Paula M. Meek, RN, PhD, FAAN¹, Terrah F. Akard, PhD, RN, CPNP, FAAN², Ardith Z. Doorenbos, RN, PhD, FAAN³, Lisa M. Schilling, MD, MSPH¹

¹University of Colorado, Aurora, CO; ²Vanderbilt University, Nashville, TN; ³University of Illinois, Chicago, IL

Introduction

Sentiment Analysis (SA) methodologies are being investigated with increasing frequency for potential applications in healthcare. One concern is that these methodologies are lacking in the ability to analyze healthcare text because their lexicons were not curated for this purpose and do not include a comprehensive set of health-related terms. The purpose of this observational descriptive study is to perform secondary analysis of previously collected patient illness narratives² to assess the feasibility and face validity of SA methods to characterize the sentiment of the illness narratives and explore how various SA methods and their lexicons impact the results. The patients’ narratives are based on patient interviews conducted by one of the authors (HC) and focus on the psychosocial and spiritual impact of illnesses, written from the perspectives of patients with serious illness.

Methods

We analyzed de-identified two- to four-page narratives (n=10) at various levels of text analysis (document, paragraph, sentence), using four SA approaches: TextBlob Pattern (TB-P), TextBlob Naïve Bayes (TB-NB), NLTK Vader (NV), and SentiWordNet (SWN). We compared the sentiment polarity (a numeric representation of the sentiment between -1 and +1), the terms used for calculating the score, and the inter-rater agreement. We also identified medical-related terms from the narratives using eight medical terminology standards (i.e., RXNORM, SNOMED-CT, LOINC, CPT, ICD9CM, ICD10CM, MESH and NCIT) from the National Center for Biomedical Ontology's BioPortal and compared them with the terms used by the SA methods, with the goal to identifying overlaps and gaps of terms.

Results

At the document level, most of the patient-illness narratives were classified as positive by all the SA approaches. However, we found more variations and disagreements among the SA approaches when polarity was evaluated at the paragraph and sentence level. The average Cohen’s kappa ranged from 0.8 to 1.0 at document level, from 0.06 to 0.49 at paragraph level and from 0.10 to 0.35 at sentence level. The most frequent parts of speech per narrative were verbs (average 18.47%) and nouns (average 17.71%) with fewer adjectives (average 5.36%) and adverbs (average 6.76%). The overall proportion of terms from the ten narratives with polarity zero is 25% for TB-P and 27.5% for SWN. TB-NB and NV do not provide lists of terms used for calculating polarity. The selected BioPortal ontologies identified an average of 27.6 (15.6%) unique terms per document that were not identified by both the TB-P and the SWN. The NCIT ontology identified the most terms from the narratives followed by the SNOMED-CT and the LOINC ontologies with averages of 144.4, 88.8 and 72.7 terms per narrative, respectively.

Discussion

The variation in the results of the four SA approaches indicates that these SA approaches have limitations and are not suited for use with patient-illness narratives. Overall, the SA methods demonstrated a lack of polarity concordance and missed potentially important healthcare terms. SA in healthcare must also consider factors related to the patient’s health status, conditions, diagnosis, and more. Our analysis of the patient-illness narratives at the sentence, paragraph, and document levels to explore the variation in the results confirmed that an assessment of the whole document might obfuscate important and meaningful variations. As we moved from document to paragraph to sentence, the variations in polarity scores increased and agreement between the SA approaches decreased. BioPortal served as a complementary evaluation to non-healthcare related SA approaches.

References

An Informatics Governance Model for a Shared Electronic Disease Surveillance System in Public Health: Minnesota Perspective

Sarah Solarz, MPH¹, Ann Kayser, MPH¹, Sarah Boneske, MPH¹, Hanna Chakoian, BA¹, Sripriya Rajamani, MBBS, PhD, MPH, FAMIA¹,²,³, Emily Emerson, BS¹

¹Minnesota Department of Health, St. Paul, MN; ²School of Nursing, ³Institute for Health Informatics, University of Minnesota, Minneapolis, MN

Introduction

The Minnesota Electronic Disease Surveillance System (MEDSS)¹ supports infectious disease surveillance at a state level for Minnesota. MEDSS was implemented in 2008 and is used for case management, contact tracing and outbreak investigations. Conditions that are reportable to public health and relevant electronic lab reports are stored in MEDSS. Its scope has expanded to include non-infectious diseases, such as hearing loss and birth defects. It is a person-centric surveillance system with ~1,494,508 events across infectious diseases, lead and community and family health programs. MEDSS has a user base across state programs, local public health agencies and external clinical partners with varying access/roles. The central information technology agency for the state and the system vendor collaborate with the MEDSS informatics team to support the system. Given this matrix of users, technical assistance, funding and priorities, it is essential to have an informatics governance model in place for oversight and ensuring smooth operations.

Methods

A systematic approach was pursued over the past few years to have various MEDSS groups facilitate the process of deciding and approving system improvements and prioritizing the on-boarding of data exchange partners. These groups needed to have broad representation to ensure various stakeholders are included and also to assure that funding requirements and user expectations were met in a timely fashion. The coordination of these groups and overall collaboration across public health and technical professionals is under the purview of the MEDSS informatics team.

Results

The MEDSS informatics team (n=5) leads the overall operations guided by user input and grant objectives. MEDSS Steering Committee (n=17) with MEDSS and ELR user groups (n=85) are integral to identifying system issues, enhancements and priorities. Technical issues are addressed by system admin team (n=5) and messaging team (n=5). MEDSS leadership (n=4) oversees all these groups. They collaborate with MDH Office of Health IT which sets e-Health policy goals, and other internal/external partners.

Conclusions

The current governance model supported by various experts has placed MEDSS in a stronger position to respond to various federal/local initiatives and also to be strategic on decisions given various constraints, including funding. Information systems across public health agencies have varying capabilities, technical support and different governance structures. The proposed model can be adopted/adapted by other programs or in other jurisdictions.

Acknowledgement: The authors thank the numerous experts who participate in various MEDSS governance groups.

References

Why is Data Harmonization Difficult?
Towards a Taxonomy for Structuring the Mapping of Variables between Data Models
Christian Sommerhalder MD1, Karel Capek MD1, Xuyang Liang MS1, Andrzej Kudlicki PhD2,
Celeste C. Finnerty PhD1,2, Suresh K. Bhavnani PhD2
1Dept. of Surgery, 2Instit. for Translational Sciences, Univ. of Texas Medical Branch, Galveston TX

Abstract
To enable high-powered multi-site studies, numerous healthcare organizations are attempting to harmonize their electronic health records (EHRs) with common data models such as PCORnet and ACT. However, the harmonization process is known to be difficult, requiring extensive manual curation by domain experts. Here we report on a case study of harmonizing data from an EHR to PCORNet, which enabled an examination of the complexity involved in data harmonization. The results led to a mapping taxonomy useful for structuring the mapping process, and for the development of scripts to populate databases.

Introduction
Motivated by the need to conduct high-powered and generalizable studies across multiple sites, numerous healthcare organizations have invested in harmonizing patient data in their EHRs with national data standards such as PCORnet and ACT. Unfortunately, EHRs in specific healthcare organizations have been designed with locally-determined conventions for naming (e.g., sex vs. gender), encoding (e.g., Hispanic vs Hispanic-Black), and data entry (e.g., free text vs. structured field). This complexity requires domain experts to manually determine the correct mapping. Here we examine the mappings used by a team of domain experts in harmonizing the data model used by the University of Texas Medical Branch (UTMB) hospital’s EHR, with the PCORnet2 data model, for building a multi-site registry to analyze opioid dependency in trauma patients.

Method
Our method consisted of the following steps: (1) assembled a team of domain experts consisting of 2 physicians, a researcher, and a database expert; (2) identified 3 research questions to analyze the association of patient variables to opioid dependency (e.g., How are trauma characteristics (type, site) associated with opioid dependency (>6 months post-injury)?); (3) identified variable names and their encodings in the UTMB data model; (4) mapped the selected variable names and encodings from the UTMB data model to the PCORnet data model; and (5) identified the types of data mappings used by the team.

Results and Discussion
To address the 3 research questions related to opioid dependency, the team identified 178 variables, of which 6 were specific to trauma (e.g., trauma site), while the remaining 172 were generic (e.g., demographics, and medications). By examining the UTMB and PCORnet data models, the team mapped the above variables to 13 PCORnet-defined domains (the number of UTMB variables in each domain are shown in parentheses): DEMOGRAPHIC (7), VITAL (13), ENCOUNTER (14), CONDITION (26), DIAGNOSIS (13), PROCEDURE (13), LAB_RESULT_CM (28), OBS_CLIN (16), PRESCRIBING (20), DISPENSING (11), MED_ADMIN (8), DEATH (4), DEATH_CAUSE (5). The following 6 PCORnet domains were not relevant to the selected characteristics: PROC_CM, PROVIDER, OBS_GEN, HARVEST, PCORNET_TRIAL, and ENROLLMENT. Furthermore, 37 UTMB variables necessitated an examination of actual patient records to determine their correct semantic mappings, which required IRB approval that is currently pending.

Examination of the above mapping process helped to identify 5 types of mapping: (1) one-to-one (e.g., Patient ID number mapped to PATID); (2) one-to-many (e.g., Drug-Screen containing amount of opiates mapped to multiple LOINC Codes to record different amounts of opiate such as 91052-1, 91043-0, 10369-7); (3) many-to-one (e.g., Operation Date and Operation Time merged into PX_DATE using the DD/MM/YY HH:MM format); (4) one-to-none (e.g., Procedure End Date, and Procedure End Time do not have corresponding variables in PCORnet); (5) derived (e.g., Sequential Organ Failure Assessment (SOFA) Score calculated based on Systolic Blood Pressure, Serum Creatinine, Serum Bilirubin, Serum Glucose, Platelets, PaO2/FiO2 Ratio, Glasgow Coma Scale). Furthermore, the above taxonomy was also useful for categorizing the variable encodings (e.g., Black-Hispanic and Non-Black Hispanics, merged to Hispanics and therefore categorized as many-to-one). We are currently evaluating whether this taxonomy covers the types of mappings required by the ACT model.

The above mapping taxonomy should be useful to help collaborating sites use a consistent nomenclature when mapping each variable, enabling more efficient communication, consensus, and reporting across sites. Furthermore, the above taxonomy enables the use of modular programming for reducing complexity and redundancy when creating extract-transform-load (ETL) scripts at each site, with the goal of more efficiently and accurately populating the registry with actual patient data.

Acknowledgements. Funded in part by a supplement to the UTMB CTSA grant (UL1 TR001439), and NIH T32-GM8256.

References
Personalized Treatment For Type 2 Diabetes Using Weighted K-nearest Neighbors

Wenyu Song, Ph.D.1,2*, Linying Zhang, M.S.3*, Emily Gill, M.B.,ChB. 4, Jeremiah Zhe Liu, M.S.1, Adam Wright, Ph.D.1,2,5

1Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, 2Department of Bioinformatics Informatics, Harvard Medical School, 3Department of Biostatistics, Harvard T.H. Chan School of Public Health, 4General Practice and Primary Health Care, University of Auckland, New Zealand, 5Partners eCare, Partners HealthCare, Boston

* these authors contributed equally to this work

Overview: In this study, we trained a weighted k-Nearest Neighbor (KNN) on real patient clinical record to prescribe medication that has the lowest predicted HbA1c for each patient visit. Our cohort consists of 6,218 T2DM patients from Partners Biobank with 42 unique T2DM medications. We grouped medications into 8 clinically distinct categories ranging from mono-therapy to combination therapy. We also extracted multiple clinically informative features, including sex, age, HbA1c and BMI measurements. Feature weights were determined for each treatment by regressing post-treatment HbA1c against all features. During the model training, for each patient visit, the algorithm found k most similar patient visits based on weighted Euclidean distance under each treatment regime. Then mean HbA1c of those k patient visits under each treatment regime were calculated and the regime with the lowest mean HbA1C was recommended by the algorithm. The algorithm was also validated in a testing dataset and similar performances were obtained.

Conclusion: Among 41,758 patient visits, the predicted mean post-treatment HbA1c under our prescriptive algorithm was lower than that of standard care by 0.24 ± 0.01% (P < 0.001) (Table1). The patient visits with HbA1c ≥7.5% benefited the most from the predicted treatment plan, where post-treatment HbA1c was 0.37% lower than standard care. All the other subgroups also had significant reduction in predicted post-treatment HbA1c. For patients visits under some form of insulin treatment, our algorithm found that prescribing both insulin and metformin could more effectively reduce HbA1c. Our algorithm also confirmed that for patients under some form of insulin therapy, switching to treatment without insulin on average led to a suboptimal glycemic control. For patients visits without insulin, metformin was predicted to be the most effective in reducing HbA1c.

In Figure1, each cell in the figure represents patients for whom the prescriptive algorithm recommended switching from the regimen on the vertical axis to the regimen on the horizontal axis, labeled with the number of patients who made that switch. (METO: metformin alone; INSO: insulin alone; OTHO: other drug alone; METINSO: metforimin+insulin; MET1: metforimin+other(1); INS1: insulin+other drug(1); METINS1: metforimin+insulin+other drug(1); MULTI: other drugs (>=2))

Table1. Predicted mean relative to standard care (SE)

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>Mean Percent</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>All visits</td>
<td>-0.24</td>
<td>0.01</td>
</tr>
<tr>
<td>Pre-insulin visits</td>
<td>-0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>Post-insulin visits</td>
<td>-0.35</td>
<td>0.02</td>
</tr>
<tr>
<td>Age&lt;60</td>
<td>-0.30</td>
<td>0.01</td>
</tr>
<tr>
<td>Age&gt;=60</td>
<td>-0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>Non-obese (BMI&lt;30)</td>
<td>-0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>Obese (BMI&gt;=30)</td>
<td>-0.24</td>
<td>0.02</td>
</tr>
<tr>
<td>HbA1c&lt;7.5</td>
<td>-0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>HbA1c&gt;=7.5</td>
<td>-0.37</td>
<td>0.02</td>
</tr>
</tbody>
</table>

All p<0.001

Figure1.

Acknowledgements: The authors would like to acknowledge contributions of Partners HealthCare Biobank for all the support. This work was supported by United States National Library of Medicine grant T15LM007092 and the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number R01HL122225.
Single-FILE: A Treatment for Multiple Digital Identity Disorder

Spencer L. SooHoo, PhD1, Anil Goud, MD1, Benjamin A. Robbins1, Pamela Roberts, PhD1, Lyna Truong1, Harold Moyse1, Shao-Chi Huang1, Lilith Huang1, Donaldo T. Rodriguez1, Richard R. Riggs, MD1

1Cedars-Sinai Health System, Los Angeles, CA

Introduction

In 2015, the National Institute of Standards and Technology (NIST), in partnership with the Office of National Coordinator (ONC), provided a grant to Cedars-Sinai to use standards-based technologies to pilot the creation and use of federated identities in healthcare to access multiple distinct Health Provider Organization (HPO) Electronic Health Record (EHR) systems. We developed Single Federated Identity Login to EHRs (Single-FILE), which supports an identity ecosystem in healthcare through federated identity broker services. Key attributes are that the federated identity can be pseudonymous (i.e. allow use of a social identity such as Facebook, Google or other identity), provide single sign-on (SSO), support multi-factor authentication (MFA), and include privacy considerations in the architecture from the onset.

Design and Implementation

Single-FILE has a hub and spoke architecture that provides identity, authentication, and authorization broker services. Until EHRs include this functionality, it requires deployment of an identity server (IdS) at each HPO that is delivered as a Docker container. A Single-FILE user selects a social credential (Google or Facebook) or one issued by Single-FILE to be used as a federated identity for SSO. The user provides a partial Social Security Number, zip code, and data of birth that Single-FILE uses to query the EHR system for a match. Once a match has been found, a One-Time Passcode (OTP) is sent to the phone number registered in the EHR. Upon acknowledgment of the OTP the federated identity is bound to the user’s EHR identity. This process is repeated for each EHR. An important component of Single-FILE is a Trust Framework, which is critical to the acceptance of a federated identity. This framework, a legal, multi-party agreement, specifies the required roles, responsibilities, legal requirements, and conformance standards for each participating HPO.

Identity proofing assurance levels (IAL) are essential when using a federated identity. NIST IAL standards1 for non-healthcare specific domains range from IAL-1 (least assuring) to IAL-3 (most assuring). IAL-2 requires resolution, validation, and verification of government issued photo identifications with validation against the issuer’s database. Since this is not feasible in healthcare operations, we proposed the use of historical transaction data through the trusted referee mechanism provided by NIST IAL standards to attain IAL-2 status.

Single-FILE allows users to perform an identity verification step via an OTP to bind a digital identity with an IAL-2 proofed identity from an HPO. This has been deployed at Cedars-Sinai and the California Rehabilitation Institute as a proof-of-concept and will allow patients to access EHR data at both HPOs using a single federated identity.

Conclusion

Single-FILE’s identity ecosystem model allows HPOs to serve as a “trusted referee” to confer IAL-2 status to user digital identities they have created. Another potential benefit is to provide a monitoring system that could help detect fraudulent activity and improve awareness to the patient of potential misuse. During the development of Single-FILE, Apple’s Health App and Epic’s Happy Together were released. These provide partial solutions for patients with multiple HPO accounts, but still lack key features of a true federated identity that can decrease both the maintenance burden and the security/privacy risk for users.

References

Detection of Opioid Overdose Cases from Electronic Health Records: A Pilot Study using Data Collected from the MIMIC III Database

Alfred Sorbello, DO, MPH\textsuperscript{1}, Rashed Hasan, PhD\textsuperscript{1}, Henry Francis, MD\textsuperscript{1}, Mitra Ahadpour, MD\textsuperscript{1}

\textsuperscript{1}US Food and Drug Administration, Silver Spring, MD, USA

Introduction:
In this pilot study, we assess three traditional approaches to detect opioid overdose cases from information recorded in the Medical Information Mart for Intensive Care III (MIMIC III) electronic health record (EHR) database\textsuperscript{1}.

Methods:
The three methods assessed are: (1) International Classification of Diseases, Ninth Revision (ICD-9) code algorithms, (2) keyword searches of free text narratives using SQL queries, and (3) natural language processing (NLP) using NLTK-python library with clinically relevant keywords/phrases and negation as defined by a clinical expert. Consecutive pre-validation cases were manually assessed by a single reviewer as suspect, probable, or confirmed opioid poisoning cases based on the CDC case definition\textsuperscript{2}.

Results:
As shown in Table 1, the highest numbers of probable or confirmed cases were identified with ICD-9 codes. Keyword searches retrieved numerous pre-validation cases, but with many false positives. NLP and keyword searches detected cases not identified with ICD-9 codes.

Table 1: Opioid Overdose Cases identified from the MIMIC III Database

<table>
<thead>
<tr>
<th>Method</th>
<th>PreV cases</th>
<th>Manually Reviewed Cases</th>
<th>Opioid overdose cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total N</td>
<td>Total n (%N)</td>
<td>Suspect (n)</td>
</tr>
<tr>
<td>\geq1 ICD-9 diagnosis code for opioid poisoning and/or \geq1 E-code for accidental opioid poisoning, self-harm/suicide, or undetermined intent</td>
<td>172</td>
<td>172</td>
<td>152 (88%)</td>
</tr>
<tr>
<td>Keyword searches</td>
<td>Facets A + C</td>
<td>3,175</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>Facets B + C</td>
<td>2,550</td>
<td>175</td>
</tr>
<tr>
<td>NLP with keyword Facets A + D, other keywords\textsuperscript{*}, and negation</td>
<td>175</td>
<td>175</td>
<td>48 (27%)</td>
</tr>
</tbody>
</table>

\textsuperscript{*}ICD-9 opioid poisoning diagnosis codes (965.00, 965.01, 965.02, 965.09), E codes for accidental poisoning (E8500, E8501, E8502), undetermined intent (E980.0 - E980.5), and self-harm/suicide (E950.0-E950.4). Keyword Facet A - Opioid Drug Facet: heroin, methadone, oxycodone, oxycodin, codeine, suboxone, hydrocodone, MS Contin, hydromorphone, morphine, buprenorphine, dolophine, tramadol, ultram, tval, fentanyl, actiq, duragesic, buprenex, permocet, vicodin, loricit; Facet B - Opioid Term Facet: opiate, opioid, opiates, opioids, opioid pain meds, opioid pain medications, opioid drugs, narcotics, narcotics, pain meds, pain medications; Facet C - Overtone Facet: overdose, overdosed, excess, excessive, too much, intoxication; Facet D - Benzodiazepine Drug Facet: clonidine, klonopin, alprazolam, diazepam, valium, clonazepam, temazepam, ativan, Xanax, zolpidem, benzodiazepines, benzo, benzos, chloridiazepoxide, Librium, chlorazepate, traxene, estazolam, prosom, flurazepam, alprazolam, oxazepam, serax, Restoril, triazolam, halcinon, quazepam, doral; PreV = pre-validation; NLP = natural language processing; NA = not applicable; *other keywords/phrases: alcohol, smoking, depression, history of polysubstance abuse.

Conclusions:
When combined with ICD-9 codes, extracting information from free text in EHRs augments opioid overdose case detection. Future work explores EHR phenotyping using supervised machine learning for case ascertainment.

References

Acknowledgements: This project was supported in part by appointment to the research participation program at CDER administered by the Oak Ridge Institute for Science and Education (ORISE) for the FDA. Funding support received from the FDA/CDER/Office of Translational Sciences.

Disclaimer: The views expressed are those of the authors and do not necessarily represent the views of the US FDA or the US Government.
A structured approach to working with wearable health data: lessons learned from the University of Chicago IBD biosensor study
Philip Sossenheimer, BA; Olivia Yvellez, BA; Anoop Mayampurath, PhD; Samuel Volchenboum, MD, PhD; MS; David T Rubin MD
University of Chicago, Chicago, Illinois

Problem:
In recent years there has been increased interest in the use of commercially-available wearable devices as a means of passive patient monitoring in inflammatory bowel disease (IBD). These devices can collect and store patient data, including steps, heart rate (HR), and sleep metrics that could be useful for passively monitoring patients. Previous studies using wearable devices suggest that patients with IBD have decreased physical activity and poorer sleep quality compared to healthy controls, however data collected from wearable devices have not successfully been leveraged to predict disease status in IBD patients. Here we report preliminary results and lessons learned from our study investigating the utility of wearable health devices in monitoring IBD.

Methods:
Patients in the IBD clinic at the University of Chicago were enrolled in an IRB-approved observational study. Each participant was given a Fitbit (Charge or Alta HR, San Francisco, CA) and installed software to collect the wearable data as well as IBD-related survey questions (Litmus Health, Austin, TX). Clinical and demographic data were captured from patient medical records and stored in a Research Electronic Data Capture (REDCap) database. The predictive ability of biosensor data for disease activity was determined by comparing Fitbit metrics from the week before a disease activity assessment between patients with active or quiescent disease, and calculating the AUC for each metric. Multivariate logistic regression was performed with five-fold cross validation to assess the predictive ability of the biosensor data for disease activity.

Results:
So far 189 patients have been consented and enrolled in the study. Patients tended to be male (n = 100, 52%), white (n = 170, 90%), and young (median age = 37, IQR = 28-47). Compliance with the biosensor was 47.9% (IQR 17.0-88.0%) during the day, and 29.1% (IQR 8.4-60.7%) during the night. Our multivariate model including steps, resting heart rate, and basic clinical characteristics had an AUC of 0.72 for predicting disease activity, with an overall accuracy of 67%.

Research Challenges and Lessons Learned:
One of our main challenges was ensuring good compliance (percent of days with recorded Fitbit data) while wearing the device for an extended period. To improve compliance moving forward, we have updated our application to include a real-time data feed, allowing us to proactively contact patients who are not wearing their device by telephone. Although we have had moderate success in predicting disease activity, we believe our results could be improved by establishing baseline characteristics for individual patients during times of objective disease remission (as established by endoscopy or serum/stool markers) and evaluating how changes from that baseline correlate with changes in disease symptoms.

Conclusions:
Wearable health devices offer significant opportunities for improved patient monitoring. Researchers interested in studying the utility of these technologies should consider ways to ensure study compliance, clearly define study outcomes from the outset, and have a clear strategy for data analysis.
Toward large-scale management of clinical quality dashboard visualizations: The Performance Summary Display Ontology

Cooper Stansbury1, Colin Gross2, Veena Panicker2, BDS, Zach Landis-Lewis, PhD, MLIS2
1University of Michigan Dearborn, Dearborn, MI; 2University of Michigan, Ann Arbor, MI

Introduction

Efforts to create and maintain clinical quality dashboards and reports are often siloed, making it hard to study the general conditions under which a particular visual artifact is useful for improving clinical practice. We propose a Performance Summary Display Ontology (PSDO) as a formal computable model for describing graphical, structural, and theoretical characteristics of information visualization artifacts. We designed and built PSDO as part of a learning health system to study when, where, and how different visual artifacts can aid in quality improvement communication to healthcare professionals and teams. Our representation of visualization artifacts is intended to be broadly generalizable while accounting for the need to represent granular differences between displayed attributes such as color, spacing, and semantic content. We believe that the use of ontology to manage knowledge about information visualizations enables not only closer scrutiny of design choices as they relate to outcomes, but also the ability to facilitate information interoperability between visual artifacts in healthcare organizations.

Methods

Using Basic Formal Ontology (BFO)1 and The Open Biological and Biomedical Ontology (OBO) Foundry best-practices, we developed easily extendable design paradigms to represent information visualization with regard to domain-specific semantics. By building on the high-level representational structures described by Zhang2 and Munzner3, we defined classes representing core components of visual artifacts and their relations. Through iterative content analysis of dashboards and reports; discussion with healthcare professionals, healthcare administrators, visualization designers, and programmers; and application of visualization theories2,3, we identified universals relating to content-type, physical design, display-type, and mutable physical properties.

Results

The representation in PSDO extends OBO Foundry ontologies concerning information artifacts (IAO) and statistical visualization artifacts (STATO) to represent visual artifacts in terms of their qualities, dimensionality, type, and sets of granular components (‘marks’). We represent the qualities and emergent properties of individual marks and mark aggregation in terms of their semantic content: a) performance measures, b) entities ascribed a performance, c) performance, and d) time intervals. Importantly, by representing the semantic content of a display at various levels of granularity, we allow for management and retrieval of artifacts about a particular entity dependent on other query constraints. By representing visualizations with greater generalizability and deeper granularity in an ontology, we enable retrieval of artifacts dependent on a variety of graphical and theoretical conditions. PSDO contains the building blocks for large-scale visual artifact management without sacrificing graphical ‘fitness for use.’

Conclusion

There is a need to manage knowledge about visual artifacts, even within a single healthcare organization, such that the effects on communication of the visual artifacts can be better understood. By strengthening ontological representation of information visualizations, we believe that we can contribute to (1) better access and management of artifacts, subject to constraints at different levels of specificity, and (2) a foundation for learning health systems that enable humans and computers to leverage each other’s respective strengths in order to understand how, why, and when different visualizations are effective.

References

Patient-Centered Care Team Communication and Coordination Tools in a Pediatric Ambulatory Cancer Care Center

Daniel Stein, MD, PhD, Eugenia Kim, MS, Keri Martin, Jennifer Kim, Aaron Ostrovsy, MPH, You Chung Jr, MSN, RN-BC, CCRN-K, Melissa Zimmermann, MS, BSN, RN-BC, Tham Nguyen, MSN, Eric Schmitz, Christina Ward, MBA, Maritza Rosa, MPH, Gregory Jordan, MaryAnn Connor, MSN, RN-BC, CPIM, Jan, Julianne Ruggiero, MSN, CPNP, Rachel Bright, BSN, RN, CPHON, Jill Ackerman, MBA, Farid Boulad, MD, Peter Stetson, MD, MA
Memorial Sloan Kettering Cancer Center, New York, NY

Introduction

Interest in care coordination has increased steadily over the years both as a means of improving patient outcomes and decreasing costs. Effective communication and coordination of care has been established as one of six National Quality Strategy priorities by HHS, and care coordination initiatives have been estimated to result in $240 billion in savings by the Institute of Medicine.1,2 At Memorial Sloan Kettering (MSK), our Pediatric Ambulatory Care Center (PACC) sees approximately 100 patients/day for outpatient oncology services, and the clinical team consists of over 100 members including physicians, advanced practice providers, nurses, nursing assistants, patient care techs, care coordinators, social workers, pharmacists, and child life therapists. The PACC leadership observed challenges in care coordination in this setting, and therefore sought to improve team member identification and communication efficiency by collaborating with our Informatics teams to implement two new digital tools into the PACC workflow.

Methods

Two new tools were introduced in the PACC: 1) A secure text messaging app for clinician communication accessed via smartphones and computer workstations; 2) A custom-developed tool for care team assignments, which integrates with the electronic health record (EHR), allows staff to view and update members of a patient’s care team, and includes features such as routing messages via the secure text messaging platform. Implementation planning and go-live of these tools were completed in multiple phases. A project team was formed, and an analysis was conducted of the current PACC workflows to assess potential impact and desired integrations, after which a series of development and configuration efforts were performed and tested. Scheduled training sessions followed prior to go-live, which occurred in two phases, the first in October 2018 with the care team tool and the second in December 2018 with the secure texting app. System audit data was collected for two months after implementation to evaluate usage in terms of team assignments made, secure messages sent, and the roles of the individuals taking these actions.

Results

For each of the two months, nearly 2,000 care team assignments on patients with visits in the PACC were made, and approximately 7,000 secure messages were sent among staff associated with the PACC. Data collected from system audits demonstrated that a diverse set of staff consistently used the secure texting app and the care team tool, with nursing showing the highest activity for both tools. The functionality to send secure messages directly from the care team tool was utilized with the highest volume of directed messages being those from nurses to advanced practice providers. Preliminary analysis of user surveys showed the tools were easy to use and useful in coordinating care.

Conclusion

Preliminary feedback from the PACC teams indicate that providing a tool for secure, closed-loop communication and a mechanism to easily identify a patient’s care team members has been beneficial for team coordination, patient flow, and patient care. Further studies involving user surveys and analysis of audit and EHR data are underway to evaluate usage of the tools over time and impact on clinical workflow and patient outcomes.

References

Use of Information Visualization to Enhance HIV-Related Clinical Interactions

Samantha Stonbraker, PhD, MPH, RN1,5, Carmela Alcántara, PhD2, Silvia Amesty, MD, MPH, MSEc1, Maureen George, PhD, RN, AE-C1, Ana F. Abraído-Lanza, PhD1, Mina Halpern, MPH5, Suzanne Bakken, PhD, RN1, Rebecca Schnall, PhD, MPH, RN-BC1

1School of Nursing, 2School of Social Work, 3Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, 4College of Global Public Health, New York University, New York, NY, 5Clínica de Familia La Romana, Dominican Republic

Introduction

Information visualizations (infographics) can enhance clinical communication related to chronic conditions. The purpose of this study was to assess the feasibility of using infographics during clinic visits to improve communication and subsequently, health outcomes in persons living with Human Immunodeficiency Virus (HIV) (PLWH).

Methods

In preliminary work, content for infographics was established with a rigorous triangulation methodology; then, an iterative participatory design methodology that included PLWH, health care providers, and visualization experts was used to develop infographics.1 For the current study, we are using a pre-test post-test design to explore feasibility of infographic use during clinic visits at a primary care clinic in the Dominican Republic. Initially, two physicians were trained to use laminated, paper-based infographics to convey health information during visits, then we enrolled 50 adult PLWH who receive the intervention at baseline (3-, and 6-month visits). At each visit, primary outcomes of CD4 count and viral load are extracted from participants’ records and participants complete questionnaires at baseline, 3-, 6-, and 9- month visits that include demographics and six secondary outcomes: HIV-related knowledge, engagement with clinician, satisfaction with provider and the clinic, self-efficacy to manage HIV, adherence, and health related quality of life. Outcomes were selected based on the pathways through which communication can improve health outcomes identified by Street et al.2 and measured with scales previously validated in Spanish. Descriptive statistics of demographics were calculated and means/standard deviations established for each measure at each time point. Mean changes between scales were determined for each time point of continuous variables using paired t-tests and Chi-square and Fisher’s exact tests for categorical variables. Alpha for significance was .05 for all tests. Data collection for final follow up visits is planned for August 2019.

Results

Participants (N=50) are predominantly female (56%), with a mean age of 40.4 years (SD ± 10.6) and have been living with HIV for a mean of 6.3 years (SD ± 6.1). For preliminary analysis, mean differences in scores between baseline and 3-month visits were calculated for the 47 participants who completed both visits. The mean CD4 count decreased from baseline (377.2 ± 273.5) to 3-months (372.8 ± 246.8) as did mean viral load from baseline (89,650.8 ± 245,218.4) to 3-months (84,861.3 ± 245,422.5) though neither change in primary outcome was significant. Participants’ HIV-related knowledge statistically significantly increased from baseline (7.9 ± 2.2) to 3-months (10.0 ± 1.4) (95% CI, 1.4 – 2.7). The engagement with provider scale scores statistically significantly decreased (as desired) from baseline (14.9 ± 3.5) to 3-months (13.3 ± 1.0) (95% CI, 0.49 – 2.60). Pain intensity also statistically significantly decreased from baseline (1.3 ± 2.3) to 3-months (0.27 ± 1.2) (95% CI, 0.50 – 1.84). Remaining secondary outcome measures showed improvements in mean scores between time periods, some with statistically significant changes.

Conclusion/Discussion

Information visualization is a key informatics approach. Preliminary data show infographic use during clinic visits is a feasible method to facilitate HIV-related communication and is associated with improvement in patient outcomes.

References

Applying SNOMED CT Concepts to CPT Codes: Enabling Consistent Queries for Health Services Researchers, and Specialty-specific and Visit-specific Views of Patient Information

Norman E. Stone III MD1, Carol A. Bean PhD, Thomas J. Neal MD MPH1, Karl J. Stiller PharmD, David F. Schmidt MD3, Andrew J. Holdaway MD2, Lauren C. Thompson, PhD1

1 DoD/VA Interagency Program Office, Rosslyn, VA 2 Defense Health Agency, Falls Church, VA 3 Veterans Health Agency, Washington, DC

Introduction & Background
The American Medical Association’s (AMA) Current Procedure Terminology (CPT)1 is the primary standard for categorizing healthcare services, from outpatient office visits and telephone calls, to inpatient rounding and surgery. The Category I CPT section consists of 9,314 5-digit numeric codes and descriptions structured in a flat, 2-level, hierarchy. This structure groups related procedures along one axis of description, typically a base procedure. Although closely-related procedures appear sequentially – e.g. knee replacement procedures – other groupings of procedures are difficult to identify – e.g. all major joint replacement procedures. Some of the most important questions about the operations of health institutions require grouping procedures around these broader dimensions of relationship.

The purpose of this study is to demonstrate the utility of applying descriptive concepts from the Systematized Nomenclature of Medicine Clinical Terminology (SNOMED CT)2 to CPT codes for improved relating of CPT codes along multiple dimensions of relationship. The intent is to (a.) provide a model for improving standardization of queries by health services researchers and health system operations officers, and (b.) to enable specialty-specific and appointment-specific views of patient health data.

Materials & Methods
The 2018 AMA CPT was reviewed and a selection of orthopaedic surgery-related codes was identified. Each CPT code and long description were analyzed and matched to a single SNOMED CT concept3 (Axis 1). The CPT Code’s long description was then segmented into atomic concepts that were each mapped to SNOMED CT concepts (Axis 2). Synonyms were identified and mapped to SNOMED CT concepts (Axis 3). The long description was then reviewed and additional SNOMED CT concepts were applied, including anatomic region (Axis 4), medical specialty involved (Axis 5), location the procedure is performed, implant or graft required, etc. (Axis 6). See Figure 1.

Figure 1: Model

<table>
<thead>
<tr>
<th>Axis 1</th>
<th>SNOMED-CT Exact Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axis 2</td>
<td>Segmentation of Long Description</td>
</tr>
<tr>
<td>Axis 3</td>
<td>Synonyms</td>
</tr>
<tr>
<td>Axis 4</td>
<td>Anatomic Region</td>
</tr>
<tr>
<td>Axis 5</td>
<td>Clinical Specialty</td>
</tr>
<tr>
<td>Axis 6</td>
<td>Other Attributes</td>
</tr>
</tbody>
</table>

Results
86 CPT codes and long descriptions were analyzed. Figure 2 demonstrates an example for CPT code 27447.

Figure 2: Example

27447

Conclusion
Applying SNOMED CT concepts to CPT codes is feasible and adds useful information for health services researchers. It enables association of related CPT procedures stored in non-adjacent regions of the CPT value set. If extrapolated to other standard terminology sets, this approach can enable specialty-specific and appointment-specific views of patient health data, in turn leading to improved clinical efficiency and safety.

References
2. IHTSDO SNOMED CT 5-Step Briefing: http://www.snomed.org/snomed-ct/five-step-briefing
3. IHTSDO SNOMED CT Browser: https://browser.ihtsdotools.org/
Accelerating Self - Service Analytics and Rapid Data Exploration using a Custom Tableau Web Data Connector (WDC) to REDCap

Olga Strachna, MS1, Alexander J. Tanchoco2, Stuart M. Gardos3, Wazim Narain, PhD1, Peter D. Stetson, MD, MA1
1 Division of Health Informatics, 2 Information Systems; 3 Memorial Sloan Kettering Cancer Center, New York, NY

Introduction/Background
REDCap is becoming widely used within the healthcare domain as a tool to capture structured, abstracted data from the electronic health record for use in clinical research1. As of today, there are over 895K users spread across 3342 institutions working on 653K projects2. Similarly, Tableau analytics software is seeing an adoption increase within healthcare organizations around the world. It offers the ability to seamlessly connect and integrate over 100 native data connectors as well as generic Open Database Connectivity (ODBC) interfaces and custom developed Web Data Connectors (WDC) which utilize Application Programming Interfaces (APIs) to extract data from applications3. As healthcare organizations become more data driven, there’s an increasing need to integrate multiple sources of data to model care pathways, streamline healthcare operations, and simplify the decision-making process. Currently, the analytics cycle from data request to dashboard completion can take approximately 1 to 2 months’ time depending on the project complexity. Our goal was to develop a secure, scalable, self-service analytics framework allowing users to access patient data quickly with the ability to ask questions in a straightforward plain text format. This allows for rapid data exploration by end-users to develop hypotheses without any coding.

Methods
We developed a secure, user-friendly, custom, production ready Tableau WDC to seamlessly connect to REDCap project data with the ability to integrate data from diverse data sources such as our data warehouse, Hadoop clusters or Splunk for real time system audit log data, as examples. In addition, we developed a self-service analytics framework for publication of governed data sources and utilizing them within the Tableau Server 2019.1 new Ask Data feature, shown in Figure 1. The Ask Data feature allows for rapid discovery and hypothesis generation by using simple human friendly question format as the interface to ask a question.

Results
As depicted in Figure 2, the analyst user can seamlessly log-in to REDCap through the WDC using their network credentials via single sign-on. This eliminates the need to manage multiple passwords or the need to look up complex API keys generated by the REDCap system. The analyst user is then allowed to view only their projects and inspect the data fields available for importing. With one click, the service securely loads all the project data into a Tableau Data file where they can further integrate numerous other data sources to create one curated data set for the analysis. This mechanism reduces the manual data entry needed to be done in REDCap systems given that the data is available in structured format elsewhere. Once published to the Tableau server, the clinical or administrative end-user can explore the data visually and rapidly without the need for any coding or backend expertise on data preparation or integration.

Conclusion
With the increase in variety, volume, and variability of data available within the healthcare domain, it’s imperative to allow clinicians and researchers to have the ability to rapidly explore their patient’s data to identify patterns in clinical care operations or patient outcomes over time. Working side by side with data analytics experts during the initial discovery phase is key to formulating a well-structured dataset. Allowing users to have direct access to their requested dataset opens the door for hypothesis development and quick insights into the data. Integrating clinical research databases such as REDCap with structured EHR, operational or administrative databases creates unparalleled opportunities for precision medicine analytics projects, where EHR data is augmented with richer data abstracted into REDCap databases. Lastly but not least, having a data governance framework and working side by side with data experts is key for the success of data driven clinical operations and research projects. Employing a data governance framework, allows for generation of high quality and robust datasets accessible for data driven projects within the clinical domain while maintaining privacy and security compliance as required by HIPAA regulations. This mechanism applies to all types of projects within any clinical domains and specialties that utilize multiple sources of information to analyze data and derive insights.

References
Initial findings of embedding safety alerts in a psychiatric electronic health record system

Gillian Strudwick, RN PhD\textsuperscript{1,2}, Jane Paterson\textsuperscript{1,2}, John Fernandes\textsuperscript{1}, Kimberly Hunter\textsuperscript{1}, Sheldon Hubert\textsuperscript{1}

\textsuperscript{1}Centre for Addiction and Mental Health, Toronto, Ontario, Country, Canada; \textsuperscript{2}University of Toronto, Toronto, Ontario, Canada

Abstract

Ensuring clinicians are aware of safety risks associated with caring for individuals with a high risk for violence is required by law in some jurisdictions. This is done to allow clinicians the opportunity to prepare and plan for encounters with patients that are safe for all involved. This poster will provide the initial findings of embedding safety alerts in a psychiatric electronic health record system that communicates potential safety risks to clinicians.

Introduction

Ensuring psychiatric clinicians are aware of safety risks associated with caring for individuals with a high risk for violence and aggression is required by law in some jurisdictions. This is done to provide psychiatric clinicians the opportunity to prepare and plan for encounters with patients that are safe for all involved.\textsuperscript{1} One way of communicating potential safety risks to psychiatric clinicians is to embed alerts or ‘pop ups’ into the electronic health record system. In this case, when a clinician opens the record of a patient who has been identified as having a high risk for a safety related incident, an alert is generated for the clinician providing the necessary alert information. The purpose of this poster is to provide the initial findings of embedding safety alerts in a psychiatric electronic health record system. This poster will provide: 1) information related to the nature and number of safety alerts generated during the evaluation period, 2) perspectives of clinicians related to their behavior once viewing a safety alert, and 3) information related to the change in safety related incident reports before and after the implementation of the safety alert system.

Methods

This poster provides an overview of a study conducted at the Centre for Addiction and Mental Health in Toronto, Canada. Ethical approval was obtained. A safety alert system was developed, and embedded in the organization’s electronic health record system. Safety alerts included aggression/violence, sexual aggression, weapons, arson, as well as two non-safety related alerts which were letter of trespass and unauthorized leave of absence risk. A mixed methods evaluation was completed with data obtained from focus groups (n=6), the organization’s incident report system and the organization’s electronic health record system. A thematic analysis was conducted of the focus group transcripts, and descriptive statistics were generated for the information extracted from the organizational systems.

Results

During the 1 year evaluation period, 90 safety alerts were generated. On average, 7.5 safety alerts were initiated each month, with the majority of these being related to aggression/violence. During the same period, 3,381 unique patients were provided care at the organization. Not all clinical areas of the organization used the alerting functionality, however the clinical area with the most alerts initiated 12 over the course of the evaluation period. With regards to clinician perception of the alert functionality, it was reported that when safety alerts were present, clinicians used extra vigilance when interacting with these individuals. It was also reported that since the alerts were used sparingly, no alert fatigue had occurred. With regards to the safety related incidents for those individuals with safety alerts present, there was a decrease in the number of incident reports for most (n= 55, 61%), and an increase in the number of reports for a smaller number (n=25, 28%). Tables and graphs that depict the study results will be present on the poster. A limitation of the study is the use of incident reports, since these are voluntarily initiated.

Conclusion In conclusion, embedding safety alerts in electronic health record systems may be one strategy to support improved levels of safety in psychiatric clinical environments, however more robust and further evaluation is required.

References


1779
Mapping Unstructured Diagnoses to ICD10 using Natural Language Processing and Web-Scraping
Theresa C. Sudaria, MS, Nam Nguyen, MS
Veradigm, San Francisco, CA

Introduction
Historic problem list data from the Allscripts ambulatory EHR data set has a high volume of free-text, unstructured text. Structuring problem list data into ICD10 codes facilitates downstream analysis. However, manually mapping free-text to ICD10 codes is a labor-intensive and time-consuming process because of the number of ways a concept can be expressed in free text. Natural Language Processing (NLP) techniques have been used with similar mapping problems that need to evaluate text similarity. Web search engines are also a useful tool to map free-text. Here, we develop an automated mapping approach using natural language processing (NLP) techniques and web-scraping of a search engine.

Methods
Problem list data from deidentified Allscripts ambulatory EHR data were used that included a subset of free-text problems that were previously mapped to ICD10. The already-mapped problems were used as a training set (576,898 problems), and subset of unmapped free-handed diagnoses were used as a test set (2000 free-text problems). For preprocessing, each of the words were lowercased, with punctuation removed, and normalized using NLTK’s WordNetLemmatizer. We performed a Bing search using python packages (BeautifulSoup and requests) with the unmapped free-text problem concatenated with the string “ICD10” as the search parameters. Regular expressions were then used to find the last occurrence of ICD10 category within the first 10 web URL strings returned. The 2 most frequent ICD10 categories were retained. Next, the training set was subsetted by ICD10 category. To compare these train set descriptions to the unmapped diagnosis, Sci-kit Learn’s TFIDFVectorizer converted the descriptions into numeric arrays, based on term frequency and rarity of words, and Sci-kit Learn’s cosine similarity returned the similarity between these arrays by measuring the cosine of the angle between them. Lastly, numpy’s argmax function found the result with the highest confidence score. To speed up the process of mapping thousands of descriptions we used a python package (multiprocessing) to use half the cores in the machine.

Results
We obtained 83% accuracy on ICD10 mapping of the test set. For the items that were matched correctly, the mean confidence score was 98.8% with a standard deviation of 4.9%. For the items that were not matched correctly, the mean confidence score was 95.0% with a standard deviation of 12.1%. When we only accept imputed ICD10 codes with confidence scores of 0.9 or higher, we boost our accuracy to over 95% while retaining 86% of our mapped codes. 59% of records with confidence scores lower than 0.9 were actually mapped correctly.

Discussion
By scraping the web for the category codes before using NLP techniques, we solved the problem of finding all variations of medical acronyms and misspellings. For diagnoses that have the same acronym, it returned the most likely, common diagnosis. It also helps us know which words to emphasize. For example, when comparing “humerus fracture, initial encounter” to “humerus fracture” or “scapula fracture, initial encounter”, web scraping knows to emphasize “humerus” instead of the words, “initial encounter”.

Conclusion
This mapping tool can be effective to improve historical data in diagnosis datasets. Currently, the algorithm fails when a provider abbreviates a rare disease that shares the same acronym for a more common disease. For future work we would explore clustering methods which would categorize diagnoses, acronyms, and words commonly used by differing provider specialties to improve precision and accuracy.

References
IRB Policies Around Initial Contact for Targeted Research Recruitment at Academic Medical Centers: A Preliminary Report

Dawn E. Sugarman, PhD1,2, Alisa B. Busch, MD, MS1,2, Jeanhee A. Chung, MD2,3, P. Pearl O’Rourke, MD2,4  
1McLean Hospital, Belmont, MA; 2Harvard Medical School, Boston, MA; 3Massachusetts General Hospital, Boston, MA; 4Partners Healthcare Systems, Boston, MA

Introduction. The identification of and outreach to patients has become easier as research use of electronic health records (EHR) increases. The potential for high-volume, targeted recruitment requires a re-examination of institutional policies regarding initial contact. The goal of this research was to learn how academic medical centers (AMCs) currently manage initial recruitment contact requests from researchers; specifically, how often a clinician known to the patient had a role in that initial contact and how often a researcher (unknown to the patient) was allowed to directly contact without involving a known clinician.

Method. A brief survey was sent to Institutional Review Board (IRB) chairs at AMCs comprising the top 50% of 2018 NIH funding (N=34). The survey asked respondents to consider the different ways that initial contact with patients with either: (1) asthma or (2) depression, as determined by medical chart review, would be permitted. For respondents who indicated that researchers may contact patients directly, we explored if and how patient permission to be contacted directly was ascertained with follow-up qualitative interviews.

Results. Survey response rate was 68% (N=23; U.S. Region: 35% West, 26% South, 22% Midwest, 17% Northeast). For each scenario, respondents could select more than one response. In both scenarios, 44% (n=10) allowed researchers to contact the patient only after a clinician introduced the study to the patient (Figure 1). 48% (n=11) of the AMCs allowed researchers to contact patients directly with no involvement of a clinician for the asthma scenario; fewer allowed this type of contact for the depression scenario (n=8, 35%).

![Figure 1. Type of contact by scenario (asthma vs. depression, N=23).](image)

The AMCs that allowed direct research contact approached patient permission in a variety of ways (Table 1). Four qualitative interviews were completed with institutions that allow direct contact (36%) and revealed that approval for direct contact was determined on a study-to-study basis by the local IRB.

<table>
<thead>
<tr>
<th>Method of contact for AMCs that allowed direct contact</th>
<th>Asthma (n = 11)</th>
<th>Depression (n = 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only patients who have elected to be contacted directly (opted in) may be contacted</td>
<td>2 (18%)</td>
<td>2 (25%)</td>
</tr>
<tr>
<td>Patients are contacted directly by default, but can opt out from future direct contact</td>
<td>5 (46%)</td>
<td>4 (50%)</td>
</tr>
<tr>
<td>Patients are contacted directly and may not opt out from future direct contact</td>
<td>4 (36%)</td>
<td>2 (25%)</td>
</tr>
</tbody>
</table>

Table 1. Method of contact for AMCs that allowed direct contact

Conclusion. Although preliminary and limited by low response rate, the results of this study show that nearly half of the AMCs allowed researchers to directly contact patients without the involvement of a treating clinician; however, this number was lower for studies involving sensitive diagnoses. As EHRs make it easier for researchers to identify patients for recruitment and demand rises from the research community to make it easier to reach out to these patients, institutions will need to re-examine their policies that support recruitment efforts, while being sensitive to patient expectations and privacy.
Plausible links from soil nutrients to human diets and obesity

Walton Sumner
University of Missouri, Columbia, MO

Introduction. Malnutrition is a major global public health problem comprising two broad syndromes, undernutrition and obesity. Undernutrition generally involves gross nutrient deficits with obvious clinical manifestations. Overweight and obesity require a surplus of calories, but also often involve subtle deficiencies of other nutrients. Human malnutrition could result from soil chemistry, crop selection, processing, and malabsorption. In fact, trace element deficiencies (e.g. Mg, Zn) have become common in soils and people worldwide. Plant, fungal, and bacterial vitamin synthesis pathways include many enzymes that use trace elements as cofactors. Therefore, soil trace element deficiencies might create predictable vitamin deficiencies in crops, and malnutrition in people.

Scott-Boyer, et al, linked multiple bioinformatics databases to identify relationships between 15 human diseases and 22 enzyme cofactors (13 vitamins and 9 trace elements). Obesity-related vitamins were B2, B3, B5, and C; bipterin; glutathione; and heme. The most strongly related element was Mg: others were Ca, Fe, Fe-S, Mo, and Zn. This study used the same databases to direct crop studies exploring how soil chemical balance might link to obesity.

Methods. KEGG synthetic pathways identified enzymes for steps from non-limiting precursor molecules to obesity-related vitamins. KEGG links to the Brenda enzyme database, which identifies vitamin and metal ion cofactors. Manual review of KEGG and Brenda records permitted tabulation of each vitamin’s typical synthesis steps and cofactor requirements, and, for each distinct element, the number of element-favoring steps (many enzymes favor Mg, but can use alternative divalent cations, like Ni). Separately, amaranth was grown on experimental farm plots managed conventionally (nitrogen fertilizer; moderate Mg deficit) versus with deliberately balanced soil nutrition (including adequate Mg). Seeds were analyzed for trace elements; vitamins B2, B3, and B5; and available calories.

Results. Mg is a favored cofactor in one or more steps in the synthesis of 7 of the 8 obesity-related vitamins (heme synthesis allows either Mg or Zn at one step). Three syntheses require Zn. Vitamin B6 is required for B5 and heme synthesis, but it has no definite dependencies. Amaranth seeds grown on nutrient balanced soils had higher K, Mg, and Mo compared to seeds grown on conventionally managed soil; lower Cu, Mn, Ni, and Zn; and nearly identical caloric density. Trends for higher B2, B3, and B5 synthesis in balanced soil were not statistically significant.

Conclusion. Biosynthetic pathway data suggested that crop content of 7 obesity-related vitamins could follow Mg in soil, and 3 could follow Zn. However, farm soil chemistry affected trace element content of amaranth seeds more than B vitamin content. Trace element deficiencies in farm soils could be causally related to obesity by one or more of at least three mechanisms: (1) Element deficiency (and perhaps related vitamin deficiencies) could impair function of obesity-related gene products; (2) “Hidden hunger” for essential nutrients (elements, vitamins, or essential amino and fatty acids) could drive excess calorie consumption while impairing functions dependent on the deficient nutrient; (3) Increased consumption of one or more alternative cations could have a toxic effect, e.g. Cu or Ni interfering with mitochondrial function. Food processing may further add and subtract nutrients, and increase calories, potentially relieving or exacerbating soil-related imbalances. Similar analyses may suggest soil connections for other diseases and guide the rational design of disease-specific diets and nutrient supplements.

References
Refinement of Health Technology Tools Through Usability Studies

Teresa Taft, PhD¹, Chuck Norlin, MD², Heidi Kramer, PhD¹, Charlene Weir, RN, PhD¹
¹University of Utah, Salt Lake City, UT; ²University of Utah Health Sciences Center, Salt Lake City, UT

Introduction
Adoption of Health Information technology tools developed to support clinician’s information needs at the point of care is often low.¹ Clinicians caring for chronically ill, complex pediatric patients often have unmet information needs. The Medical Home Portal² (MHP) online pediatric medical home reference guide was created to serve clinician’s information needs, but has experienced limited usage since its deployment in 2008. The goal of this work was to employ usability assessment methods to identify issues and design solutions aimed at increasing adoption of the tool.

Methods
A 3-tiered usability evaluation was conducted: 1) heuristic assessment, 2) user observation; and 3) expert review. First the investigators conducted an evaluation of the MHP based upon Nielsen’s design heuristics.³ Second, investigators observed 6 pediatricians interacting with the MHP using a complex pediatric vignette. Each physician generated 3-5 of their own information needs in caring for the patient. “Think- aloud” protocols were used to record the user’s perspectives as they searched for the information in the MHP. Search tasks and usability issues were identified based on pediatrician comments and heuristic findings were confirmed. Finally, investigators and designers developed technically feasible solutions that were evaluated by 4 pediatrician experts, to focus improvements to the tool.

Results
Studies demonstrated that information pediatricians sought was available in the MHP but was often difficult to locate. A sample of targets, search strategies, issues with completing the task, and design solutions are described in Table 1.

Table 1. Targets, Search Strategies, Issues, and Design Solutions

<table>
<thead>
<tr>
<th>Target</th>
<th>Search Strategies</th>
<th>Issues</th>
<th>Design Solutions</th>
</tr>
</thead>
</table>
| Diagnosis module | - Typing diagnosis in main search bar.  
- Clicking 1st, 2nd, or 3rd header tab. | Some search result labels were misleading | Replace the search tool. |
| Information related to the diagnosis module | - Scrolling & scanning the diagnosis module page.  
- Clicking items in the diagnosis module side menu.  
- Clicking external links.  
- Clicking a tab in main header navigation bar.  
- Typing terms in page search bar. | Navigation tabs disappeared with scrolling.  
Users did not open nested lists.  
Links to some websites were not active.  
Hyperlinks within diagnostic modules sometimes led to information that was not specific to the diagnosis. | Correct the error in code.  
Remove nesting of lists by default.  
Allow user to nest.  
Regularly check for broken links and repair or remove them.  
Remove general links from diagnosis modules. |
| Information for caregivers | - Users searched through most areas of the website in looking for parent/caregiver information. | Search mechanisms varied across pages.  
Half of main header tabs clearly conveyed contents; half did not.  
‘Parent & Family’ pages were lengthy and written at a ‘very high’ reading level. | Adopt a uniform search display and method across all pages of the website  
Refine main header tab labels and content groupings.  
Refine ‘Parent & Family’ pages to reach a 6th – 8th grade reading level. |

Conclusion
Heuristic evaluation, a user study, and expert review were effectively combined to identify issues and design solutions for iterative refinement of the underutilized Medical Home Portal. This process may be useful in optimizing other underutilized medical information tools.

References
2. Department of Pediatrics, University of Utah. Medical Home Portal. Available online at: www.medicalhomeportal.org
Privacy-preserving and non-intrusive video recognition technique to count the number of patient room visits during patient deterioration: a simulation study

Ahmad P. Tafti, PhD, Edin Cubro, Hongfang Liu, PhD, Brian W. Pickering, MB, BCh, Vitaly Herasevich, MD, PhD
1Division of Digital Health Sciences, Mayo Clinic, MN
2Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, MN

Introduction

Video analytics and image recognition are on the rise outside of clinical practice, in many applications ranging from self-driving cars and remote sensing to robotics and augmented reality. Even as artificial intelligence and video recognition continue to evolve, we still do not see much of these technologies implemented at the bedside in healthcare. Using currently available frameworks for automatic video understanding, our broad objective is to extend computer vision to the clinical practice to shorten the time to and increase the reliability of recognition of deterioration in hospitalized patients. It is anticipated that computer vision applications will contribute data to optimize the performance of predication analytics. In addition data gathered may be used to augment EMR patient data that is intermittent and delayed in that setting. As a first step to construct and evaluate augmented computer systems, we performed a simulated experiment using an overhead door camera combined with a computer vision algorithm to test the hypothesis that this approach could reliably count the number of visits to patient’s room.

Method

Counting frequency of visitation at any given time, can be estimated based on the two major computer vision-based strategies: 1) The line of interest (LOI) where a camera is fixed on a virtual line, thus people crossing this line are counted; 2) The region of interest (ROI), in which people in certain regions are counted. In this contribution and for the sake of flexibility and generalizability to different environmental settings, we used the ROI-based method. The proposed pipeline of the system along with a sample input and output of the system are shown in the figure.

Result

For the experimental setup, we used three types of video cameras simultaneously, Sony RX0 (1080p), Sony Action Cam HDR-AS100V (720p), and Sony Action Cam FDR-X3000 (4Kp). We also captured the video streams on both 60fps and 30fps bases. To validate the robustness of the method, besides illumination changes, we asked our volunteers to provide some sort of challenges (e.g., stay at the gate, move their hands across the gate, etc.). Comparing the ground truth visitation frequency with those estimated by the algorithm, we can summarize the results as: 1) Using the 60fps basis, we achieved accuracy of 91.31% when the room light was on (470Lux) and 86.96% when the room light turned off (145Lux); 2) 4Kp and 1080p versus 720p resolutions did not change the results, where 720p speeds up the process and it needs less storage devices; 3) Using the 30fps video feed, we obtained accuracy of 82.61% when the room light was on and 78.27% when the room light turned off.

Discussion and Outlook

We have shown that it is feasible to implement privacy-preserving (using overhead camera only) and non-intrusive computational strategy (using RGB camera only) to reliably count the number of visitation to a patients room. This work opens research avenues for further qualitative and quantitative analysis to more efficiently identify patterns of visitation frequency to patients’ rooms, with anticipated applications including better healthcare delivery and documentation. As part of our ongoing works, we are utilizing supervised machine learning methods to recognize and classify visitors role by scrubs and clothes.
Implementing a System for Predicting and Notifying When Patients are At-Risk for Acute Physiologic Decompensation

Abdul A. Tariq, PhD¹, Rohit Chaudhry, MS¹, Natalie H. Yip MD², Kui Tang, MS¹, Ryan DeCosmo, BS¹, Ruchi Shah, MS¹, Robert A. Green MD¹, David K. Vawdrey, PhD¹,³

¹Value Institute, NewYork-Presbyterian Hospital, New York, NY, USA; ²Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA; ³Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA

Introduction

Unanticipated patient decompensation affects patient outcomes and can prove fatal. In our institution, nearly 15% of all 2018 in-hospital deaths were “unanticipated” i.e. the medical team did not foresee this outcome. Early warning systems, based on patient acuity scores, have been proposed to identify such cases in advance, so that the medical team can intervene and potentially prevent adverse outcomes. In this study, we report on the development and deployment a mobile-phone-based Early Warning System (EWS) across five campuses of the NewYork-Presbyterian Hospital (NYP). Inspired by the RE-AIM framework for making informatics interventions successful, we provide best practices and lessons learned from our deployment (named “NYP EWS”) to guide other hospitals.

Methods

The NYP EWS is based on a measure of acuity from the literature¹. The system is deployed on twenty five medicine units across the enterprise. NYP EWS computes dynamic patient acuity scores based on real-time patient vitals from flowsheets. When the score crosses a critical threshold, the system automatically sends a mobile phone notification to the patient’s care team, enabling them to respond as appropriate. To minimize alert burden, we silence the system for four hours after any alert. As such, the majority of residents and nursing receive fewer than 3 alerts per day (Fig A; B).

Results

Since go-live in October 2018, NYP EWS has issued more than 4000 alerts on nearly 15% of the population of the units on which it is active. A governance committee, comprised of clinical and IT leaders, oversees enhancements to the system based on regular feedback sessions with nursing and physician partners. For example, in response to resident complaints about over-alerting, we discovered that while the alert burden was similar for residents and nurses, the physicians were alerted in clustered bursts versus an even distribution for nurses (Fig C; D; each point reflects alerts/provider per day). Relaying these results back to the physicians qualitatively improved system perception.

Conclusion

We believe that our system is trusted and utilized because we developed it in partnership with stakeholders, tailored the application to minimize burden on the clinical staff, and continuously iterate on the design to incorporate feedback. We believe that these are useful principles to guide such early warning system deployments elsewhere.

References

Identifying Non-VA Care Encounters using ADT Messages: Early Data from Two VA Medical Centers

Will L. Tarver¹, Omar Bouhaddou,² Nitin Jain,² & Brian E. Dixon¹,³,⁴
¹VA HSR&D Center for Health Information & Communication, Roudebush VA Medical Center, Indianapolis, IN; ²US Department of Veterans Affairs, Washington, DC; ³Indiana University Fairbanks School of Public Health, Department of Epidemiology, Indianapolis, IN; ⁴Regenstrief Institute, Center for Biomedical Informatics, Indianapolis, IN;

Introduction
The Veterans Health Information Exchange (VHIE)¹ enables the exchange of data between the Veterans Administration (VA) and outside healthcare organizations. Within the VA, admission, discharge, transfer (ADT) alerts can be important notifications to support quality improvement as well as care coordination. Using VHIE, non-VA community health partners can notify VA clinicians and staff when Veterans seek care outside the VA.² This can lead to timely and appropriate follow-up after patient discharge.

Methods
In this study, we sought to describe early patterns of non-VA alerts from community partners using data extracted from ADT messages sent via VHIE. A notification protocol within the electronic medical record was configured to notify VA providers of a patient’s non-VA event. Sent within 24-72 hours of a patient’s non-VA event, these alerts include information such patient demographics, visit type (e.g., emergency, inpatient, outpatient), and changes in status (e.g., discharge from or transfer to a facility).³ We examined ADT notifications sent to two VA medical centers located in Salt Lake City, UT and San Diego, CA during the 2017 calendar year. Descriptive analyses were used to examine patterns in the data received from non-VA facilities as well as the incidence of non-VA care.

Results
A majority of Veterans seeking care in non-VA facilities were male (86.8%) and non-Hispanic white (78.6%) followed by black (2.0%). Outpatient visits constituted 64.4% of total non-VA encounters while emergency room and inpatient admissions constituted 22.2% and 13.4%, respectively, in Salt Lake City. Similarly, the San Diego site observed 53.6% of total outpatient visits while emergency room and inpatient admissions constituted 29.3% and 17.1%, respectively. In the outpatient setting, an average of 24 patients in Salt Lake City (60.6%) and 141 patients in San Diego (54.2%) were identified as repeat visitors (2 or more visits) within the same month representing the largest proportion of repeat visits.

Discussion
ADT notifications provide an important opportunity for population health management and improving the health outcomes of Veterans at high risk for hospital readmission. ADT alerts potentially offer several benefits which include identifying patients that are hospitalized or seeking care in other medical facilities, improving health care quality by increasing provider awareness of their patients’ health history, allowing hospitals to track patients and prompt intervention when needed, and supporting care coordination and transition. Lessons learned may provide solutions to challenges faced in population health management such as tracking patients between discrete points of care using HIE.

References
INTRODUCTION – In electronic health record system (EHR) optimization, resources are devoted to continued improvement in EHR user competence\textsuperscript{1,2}. Physician burnout secondary to excess time spent using the EHR is common\textsuperscript{3}. Few EHR optimization processes are described in the peer-reviewed literature. We therefore developed an institutional Epic EHR optimization program for efficient EHR use to decrease physician frustration and burnout.

METHODS - Two-hour small-group sessions led by clinical informaticists and systems analysts on efficient EHR use were offered to all MSHS outpatient practices based on practice characteristics like burnout risk and interest, via e-mail and interpersonal discussion, starting in July 2018. Curriculum elements include efficient chart review strategies and EHR customization tips. Pre- and post-optimization EHR usage data was obtained using the Divurgent reporting interface. Provider satisfaction surveys were administered to assess the effect of optimization on subjective wellness immediately pre- and post-session, and at six weeks post-session.

RESULTS - Between 6/2018 and 5/2019, 363 providers in 98 practices underwent optimization training. Table 1 lists themes that emerged during sessions. Figure 1 shows number of minutes that participating providers spent documenting in the EHR on off duty days, before and after optimization training.

DISCUSSION - Optimized providers show a trend towards less time spent documenting in the EHR on off duty days. Themes reported are informing future optimization efforts at MSHS including training of EHR champions, i.e., advanced users, who can provide long-term local support to peers. Survey results show lasting optimization effects for improved efficiency, confidence, and satisfaction; qualitative comments validate in-session themes and indicate provider desire to troubleshoot and improve specific functions within the EHR.

**Table 1 - Themes Discovered During Optimization Sessions**

- New-hire online EHR training not sufficient for daily practice
- New and current providers benefit from at-the-elbow support on EHR use
- Ongoing dissemination of EHR tips in a structured format at the individual practice level is useful
- Individual practices should have representation at institutional IT workgroups

**REFERENCES**

Introduction
The electronic health record contains a wealth of clinical information as unstructured free text. This information has the potential to enhance clinical decision support systems, define at-risk populations, or provide insight for secondary research. Unfortunately, extracting this data is difficult and time consuming. Many commercial enterprises have developed natural language processing (NLP) systems to solve this problem, however these systems may not have been trained to recognize clinical text and are often expensive. An open source project known as the Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES) was developed as an attempt to fill this gap. However, while cTAKES provides health specific NLP under a free license, the application runs as a single process and analyzing large sets of notes can take months or even years to complete.1 As a result, many institutions only run cTAKES on a small cohort of patients. This can lead to significant underutilization of important clinical data. At Children’s Hospital of Philadelphia (CHOP), we sought to develop a fault-tolerant and scalable cTAKES pipeline capable of processing hundreds of millions of clinical notes in parallel.

Methods
The cTAKES pipeline was developed for use in Arcus, the CHOP Research Institute data strategy. The Arcus database contains over 114,500,000 clinical notes with approximately 50,000 new notes added each day. The data spans a period of 19 years. We chose to focus initially on cTAKES since it is an open source Apache project and a common standard for health-related NLP. To determine the appropriate infrastructure, we evaluated multiple processing frameworks including the Unstructured Information Management Application Asynchronous Scaleout (UIMA AS), Apache Spark, and a public cloud provider’s pipeline execution framework. Each infrastructure was evaluated on its ability to provide a fault-tolerant, distributed, and scalable processing environment. The pipeline was developed in Java to align with cTAKES. Testing was performed on a small subset of the Arcus clinical note data.

Results
We developed a fault-tolerant, distributed, and scalable cTAKES pipeline using the Apache Beam programming model. We chose to use Beam given its well documented Java software development kit as well as its support for a variety of processing backends including a local Java virtual machine, Apache Spark, and a public cloud provider’s pipeline execution framework. The multiple supported backends make it easier for our pipeline to be used by other institutions. Our production pipeline runs on a public cloud provider’s processing backend, which takes advantage of the cloud platform’s network configuration and native auto scaling based on application throughput.

The pipeline was tested on 3,446,233 notes, which took 44 minutes and provided 18,846,466 ontology mappings from the Human Phenotype Ontology. Use of the cloud-based processing backend allowed our job to scale to 503 machines, each with 4 central processing units and 15 gigabytes of random-access memory. Total cost for the job was $57.23. It is estimated that running this volume of data using out-of-the-box cTAKES would take over 21 days. Our research also suggests that both UIMA AS and Apache Spark would increase performance when compared to out-of-the-box cTAKES, but each required existing infrastructure that did not dynamically scale.

Discussion
We developed an NLP pipeline capable of processing large volumes of data using open source technologies to support code reuse across institutions. The pipeline runs daily and auto scales based on application throughput. NLP systems have made great advances in creating structured representations from unstructured clinical text. The derived data can then be used for a variety of purposes including clinical decision support and providing filtering criteria for research cohorts. Unfortunately, due to the time and cost associated with implementing and running NLP systems, only a few institutions have realized the potential of NLP for very large cohorts and clinical data sets. Further development of fault-tolerant and scalable health-related NLP is required.

References
The More You Know: Displaying Individualized MME Reports During Discharge Order Entry Reduces Opioid Prescribing

Bryan C. Thorne, MD PhD, Jason A. Lyman, MD MS, Rena Morse, BA, Bethany M. Sarosiek, RN, C. Page Muthusubramanian, FNP, Traci Hedrick, MD, David Ling, MD University of Virginia Health System, Charlottesville, VA

Introduction
As part of the University of Virginia Health System's (UVAHS) efforts to address the opioid epidemic, we implemented a non-interruptive decision support tool in a subset of our surgical discharge order sets in an effort to improve opioid prescribing for post-operative patients. Baseline data suggested that opioid prescribing was not being personalized based on the needs of patient, instead relying on default configuration within the order sets. We partnered with our Enhanced Recovery After Surgery (ERAS) team to develop an evidence-based approach to post-operative discharge opioid prescription, and modified our colorectal discharge orders to display recent opioid administrations and the associated Morphine Milligram Equivalence (MME).

Decision Support Tool
The ERAS team conducted a literature review and developed recommendations around post-discharge pill quantity based on the MME amount of oral opioids taken by the patient in the 24 hours prior to discharge (Table 1). Because our EHR does not natively calculate MME for inpatient opioid administration, we developed rule-based logic to calculate a maximum MME for the previous 24 hours. This was calculated by multiplying the highest dose of each opioid given by the number of doses of that opioid, and multiplying by a standard conversion factor to get MME. These rules were developed for the three most commonly used oral opioids on the inpatient formulary at UVAHS. The 24 hour maximum MME was displayed to clinicians in the pain medication section of the ERAS Colorectal discharge order set, along with explanatory text and prescribing recommendations. An order group containing oxycodone orders with the recommended dispense numbers was created to ease adherence to the recommendations.

Results
The discharge opioid prescribing tools were released in August of 2018. We analyzed data from August 2018 – February 2019 in the post-intervention group, and compared with that same time period one year earlier. We excluded patients taking opioids prior to the surgery from our analysis, leaving 145 patients in the pre-intervention group and 196 patients in the post-intervention group. Through February 2019, the percent of patients discharged without a prescription for an opioid increased from 12 % to 21 % (p =.043). For those patients who were discharged with an opioid and for whom the order set was used, the average days supply decreased from 5.4 to 4.4 (p =.011), and the average quantity dispensed decreased from 32 pills to 26 pills (p <.001). We also monitored repeat prescriptions of opioids within 14 days of discharge as a balancing measure and found no significant differences (8% pre-intervention vs. 7% post-intervention, p=.837).

Conclusions
Using a relatively simple intervention consisting of displaying inpatient MME information to providers at the point of ordering, providing a recommendation based on this information, and orders with quantities defaulted matching the recommendation, we were able to achieve a reduction in the number of patients being discharged with opioids and in the quantities of opioids prescribed. We noted strong variation between residents as to their use of the order set and prescribing practices. This pilot was successful at decreasing the amount of opioids prescribed, and our next steps will include standardizing onboarding for the resident coming on service as well as spreading the learnings and standardized discharge prescribing practices to our other ERAS surgical programs.

References
Biological Network Knowledge in Molecular Tumor Boards

Thorsten Tiede, Dipl.-Inf. (Bioinformatics)\textsuperscript{1,2}, Oliver Kohlbacher, Prof.\textsuperscript{1,2,3,4}

\textsuperscript{1}Applied Bioinformatics, Center for Bioinformatics, University of Tübingen, Tübingen, Germany; \textsuperscript{2}Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany; \textsuperscript{3}Institute for Translational Bioinformatics, University Medical Center, Tübingen, Germany; \textsuperscript{4}Biomolecular Interactions, Max Planck Institute for Developmental Biology, Tübingen, Germany

Introduction

Molecular Tumor Boards (MTBs) are interdisciplinary meetings with the aim of suggesting optimal, personalized treatment plans for cancer patients based on molecular data (genomics, transcriptomics). The recommendations are based on molecular modes of action rooted in systems medicine and rely on prior knowledge from public databases. Web resources like cBioPortal already offer visual representations of the interaction-network context of potential cancer-driving genes that can be used in MTB discussions. Integrating omics data on the level of biological networks in an interactive (visual analytics) approach can provide additional insights and speed up the discussions in MTBs.

Method

We developed a mapping of biological network data from the Systems Biology Markup Language (SBML) to a neo4j graph database, making them searchable and annotatable via the Cypher Query Language. Based on this mapping and subsequent identifier-based merging strategies, we can easily load and persist arbitrary public biological network resources available in SBML. A RESTful interface gives access to and allows filtering on the networks by interaction type (e.g., extract only phosphorylation events) or node type (e.g., only genes or metabolites). Additionally external scalar or categorical annotations can be mapped to nodes and/or edges (e.g., patient-specific expression values, confidence scores, drug target information) and simple graph algorithms can be executed (e.g., shortest path, breadth-first and depth-first search to extract network contexts, de novo pathway enrichment). After each step a new network-representation is persisted in the database, which can be used in subsequent steps. In this way intermediate results are not lost and can be reused, allowing for easy interactive exploration of the data. Implementation of the W3C Provenance Model makes backtracking of the steps taken to create any network in the service or any element therein possible. Additionally, access to arbitrary database contents can be restricted on a user basis, enabling collaborative work while still enabling data privacy, which is essential for patient-specific data. The created networks can be downloaded in the GraphML-format and may contain an arbitrary number of annotations on the nodes and edges defined by the user. Graph drawing tools like Cytoscape can then be used to visualize the graphs and detect user interaction with the network. Interactive navigation of the graphs is accomplished via subsequent calls to the service.

Results

The unified graph for protein-protein interactions derived from 221 non-metabolic KEGG pathways consists of 4,311 nodes and 73,586 undirected edges capturing interaction events even across traditional pathway-boundaries between genes and gene-products. For network contexts generated from this graph with a reasonable amount of nodes n and relationships r (n \leq 100, r \leq 500), sub-second response times can be achieved, enabling interactive analysis of data within graphical user interfaces for MTBs. A caching middleware allows to precompute time- and resource-intensive analysis steps in a preparation phase making them readily available in a presentation phase. The modular service-oriented design-approach using REST interfaces makes it easy to integrate in any existing clinical infrastructure and thus available to MTB processes. The open source software is available on https://github.com/kohlbacherlab.

Conclusion

We present a service for generating, annotating and analyzing biological networks in a clinical setting to help elucidate patient-specific disease-driving forces. We exemplify the viability of this approach by providing a mechanism to bring personalized network-contexts to Molecular Tumor Boards via the widely adopted GraphML interchange format and a responsive and modular RESTful application.
From User Experience to Digital Health Adoption (UXA): A Synthesized Framework to Guide Nursing Science

Dante Anthony Tolentino, MA, RN, Doctoral Candidate1, Sheila M. Gephart, PhD, RN1
1The University of Arizona College of Nursing, Tucson, AZ

Introduction: Many digital health technologies, including the Electronic Health Record (EHR) originate from a mechanistic isolationist characterization of workflow. The combination of a mechanistic view and the rapid implementation of the EHR yielded unintended consequences (UCs) and negative user experiences (UX) though their description is underdeveloped in nursing. UCs include technology-induced errors, unfavorable workarounds, additional work, and technology overdependence that hamper situational awareness and impact decision-making.

Purpose: To describe the synthesis of concepts from Task-Technology Fit theory and the Technology Acceptance Model to derive a new, synthesized framework that addresses antecedents and outcomes of nurses’ UX.

Method: The User Experience to Adoption (UXA) conceptual framework was formulated using Walker and Avant’s Theory Synthesis method. It builds on empirical evidence from nursing studies on UX from 2008-2018. This framework is a byproduct of the initial evidence from a comprehensive literature review. We pulled together available information about nurses’ UX with the EHR and organized the different concepts and statements to represent user experience and technology adoption.

Results: In the UXA Framework (Figure 1), UX is a precursor to actual system use, which is aligned with the construct of adoption. Nurses’ actual system use is derived from the fit among task, technology, individual, and organizational characteristics. This interplay impacts UX consequences, which are categorized across functional, physical, perceptual, cognitive, psychological, and social dimensions. In this framework, actual system use is mediated by user intention and influenced by UX consequences. Task characteristics include documentation demands, non-functional data entry requirements, task switching, regulatory impact, and information access. Technology characteristics include hardware, software, and environmental design factors. Individual characteristics encompass computer attitudes, skills, literacy, age, clinical experience, training, and previous experience with EHR. Organizational characteristics include users’ input to design, authentic leadership, organizational support and communication, work setting, and work shifts. The model depicts the multi-dimensional outcomes of UX encompassing interruptions, performance, satisfaction, workload, workarounds, documentation time, and communication processes. We will test the framework by using quantitative measures. We will use data mining to describe nurses’ task characteristics as measured by nurses’ EHR navigation patterns. We will also use regression analysis to quantify the relationship between the antecedents and outcomes of user experience. We will measure technology characteristics by the preferences and settings for individual nurses’ roles. We will assess individual characteristics by nurses’ demographic information. We will measure organizational characteristics using the nurses’ hospital site. Finally, only the cognitive dimension will be measured at this time using time spent in the EHR and cognitive workload using NASA-TLX.

Conclusion: The UXA framework can be used in a wide variety of practical applications including describing the antecedents and consequences of UX with the EHR. It can be used to design interventions to prevent unintended consequences and to identify factors for assessing meaningful use and adoption of EHR. The framework may have applicability to other technologies beyond the use of the EHR.

References:

Acknowledgements: Dr. Gephart received support from AHRQ (K08HS022908, R01HS 25984-01). Content is solely the responsibility of the authors and does not necessarily represent the official views of AHRQ.
Cross-Organizational Knowledge Sharing in the Department of Veterans Affairs (VA) to Create a Central Gateway to Information About VA Data

Timothy N. Trautman, BSD1,2; Jenifer E. Stelmack, MSW3; Amy D. Caudill, BA4; James R. Peterson, BS4; Rebecca M Gibbs, BS, MS, MBA4; Maria Souden, MSI, PhD3,5

1VA Informatics and Computing Infrastructure, Salt Lake City, UT; 2University of Utah, School of Medicine, Salt Lake City, UT; 3VA Information Resource Center, Hines, IL.; 4National Data Systems, Veterans Health Administration, St. Petersburg, FL; 5University of Illinois, School of Information Sciences: The iSchool at Illinois, Champaign, IL

Abstract

In a complex healthcare environment with a wide range of data resources, understanding how cross-organizational stakeholders can bring together and disseminate information that educates researchers about existing available data is essential to conducting research. This poster presents a real-world example of how cross-organizational stakeholders at the Department of Veterans Affairs (VA) worked together to provide researchers and operational VA data users with a central gateway to information about VA data.

Problem

The Department of Veterans Affairs (VA) maintains nearly two petabytes of data related to Veteran’s health and healthcare, which are collected from over 8,000 sites and organized in multiple databases managed by multiple offices. Understanding the VA data landscape, the content, the structure, utility of available data, and processes for requesting permission to access these data is essential to conducting research in this complex environment. While the VA maintains information about these data, it is spread across multiple locations within the VA network making it challenging for clinical and health services Researchers to determine whether the data they need exists. Thus, it was conceived that an effort was needed to bring all the information into one source – the VHA Data Portal.

Project Purpose

The VHA Data Portal was born from key stakeholders’ interests in consolidating information about vital organizational data resources that support VA data users, including Researchers. A small cross-organizational Portal workgroup was charged with developing an integrated solution that would meet each office’s unique mission, requirements, and customer needs. The workgroup began by surveying information available on the websites of key stakeholders with expertise in VA data and in developing and disseminating data-related resources. The development process involved resolving differences to present information about data resources with a coherent architecture and voice.

Product

The VHA Data Portal is a collaboratively developed VA Intranet site which brought together resources from four VA offices to support a common goal of providing a central knowledgebase about VA data and data access processes. The Portal is organized into seven main categories: data sources, data access request processes, tools and applications, resources (such as data quality and technical reports, user guides, data documentation), training and education, policy and administration, and user support. It now integrates information about VA data from multiple VA offices into a single location and serves as a central gateway to this data knowledge.

Conclusion

A central gateway to data knowledge is essential for educating Researchers and operational data users about available data in a complex healthcare environment with a wide range of data resources. As measured by user feedback, frequency of use, and interest from other program offices seeking to contribute, the VHA Data Portal has had tremendous success in consolidating data information that serves the needs of all data users and supports research that improves Veteran healthcare.

Learning Objective

Our objective was to understand the problems that Researchers and operational data users face in finding, understanding, and using data and how a solution of consolidating data information can save immense time and frustration leaving Researchers more time to conduct research with better outcomes due to a better understanding of available data and its characteristics. Operational data users experience a more streamlined process to gather the data they need to determine more profound performance measures that will lead to greater operational excellence within the VHA.
Addressing the Challenges of Telemedicine Implementation Using a Critical Success Model Approach

Tulu Bengisu¹, Cynthia LeRouge², Nelson King³, Spurthy Dharanikota⁴

¹Worcester Polytechnic Institute, Boston, MA; ²Florida International University, Miami, FL; ³Khalifa University, Abu Dhabi, UAE; ⁴Florida International University, Miami, FL

Introduction and Background

The American Heart Association recommended telestroke, a web-based application of telemedicine for acute stroke, as a low-cost and effective tool to improve access to critical stroke care and quality of care within underserved and rural regions. Understanding what is truly critical to success of telestroke, and emerging health services technology innovations, in general, is vital given competing innovation paths, constrained resources, and pressure to show promising outcomes. Critical Success Factors (CSF) is a persistent construct found in the health information technology (IT) literature that report long lists of “important things to consider” when embarking on digital initiatives. Such published lists seem to have strayed from the empirical roots, meaning, and application of this construct to discern critical knowledge, identify relationships and responsibility, and show a pathway to action. We use a longitudinal case of a telestroke implementation to revisit and rekindle a deeper analysis into the meaning of CSFs. Exploring CSFs for telestroke helps health organizations realize the value that expedited access to care via telemedicine can afford, and guide policy makers away from legislating innovations that hinge upon realizing difficult to achieve factors. We use an empirically rooted critical success factor framework (Figure 1) to explain the success factors in the form of facilitating actions. Engaging in facilitating actions increase the probability of success, and the lack of facilitating actions decrease the probability of success. The interim outcomes resulting from facilitating actions can be described as those outcomes that overcome barriers and address key implementation challenges. Our work specifically addresses: 1) What challenges do these CSFs address? 2) What success factors (key facilitators and interim outcomes) are associated with a telestroke implementation project?

Methods:

Three data collection rounds - Round 1 (prior to go live), Round 2 (go-live) and Round 3 (5-year reflective lookback) occurred during the 7-year longitudinal study in a Midwest telemedicine network with two hub hospital sites. This study used multiple methods including 27 semi-structured interviews, observation, and focus groups. Inductive and deductive coding methods identified a universe of “important factors”, grouped important factors by level, and applied an outcome-driven meaning of criticality to the resultant lists and associated quotes to discern “critical” ones. The questions in Round 3 were framed such that the interviewees revisited the factors identified as critical in the previous interview rounds and their perception about the importance of those factors was recorded. The corroborated findings were then grouped into the final list of key challenges and critical factors. The last step involved revisiting extant literature of the CSF construct to evolve a meaning and labels as well as a process for identifying CSFs that could carry across to other HIT contexts.

Results:

We identified telestroke challenges in our case for each of the areas depicted in Figure 1 – Industry Structure (e.g., limited reimbursement and specialist shortage), Strategic Context (e.g., program sustainability), Environment (e.g., limited mastery of telemedicine), Temporal (tPA guidelines of 3 hours), and Role (buy-in from stakeholders). In working down our model we identified 45 important telemedicine factors (facilitators and interim outcomes): 5 at industry level (e.g., telestroke protocols), 10 at organizational level (e.g., business plan, board of director support), 16 at sub-organization level (e.g., leverage existing infrastructure, community education), and 14 at individual level (e.g., persistent effort). Of these, 17% were determined unique to telemedicine (e.g. tele-personable clinicians).

Conclusion:

We hold that the prevalent long lists of “things to be considered”, popularized as CSFs, are iterations of best practices and perhaps a starting point, but not an end, for deep CSF analysis. The original outcome-driven meaning of CSFs holds insight into the challenges to overcome and should be revised by researchers for use and possibly empirical extension, particularly in the general area of HIT and specifically telemedicine.
Canary Library: a Sharing and Learning Community for NLP Developers

Alexander Turchin, MD, MS¹, Joonghee Kim, MD, MS¹, Sammy Elmariah, MD²,³
¹Brigham and Women's Hospital, ²Massachusetts General Hospital and ³Harvard Medical School; Boston, MA

Introduction
Narrative documents contain a large fraction of clinical data. NLP technology can help extract this information. Nevertheless, widespread adoption of NLP has yet to materialize. This is in part due to dearth of publicly available, portable and user-friendly tools. The result is two important barriers to NLP utilization: a) development of NLP tools typically requires extensive computer science expertise and b) it often involves a significant time commitment.

We sought to create a solution that would help researchers overcome these barriers. The first step was development of Canary, a GUI-based NLP-based information extraction platform that does not require software engineering or computer science expertise. It has been developed for processing narrative electronic documents to support the extraction of clinical concepts via user-defined knowledge discovery criteria and lexicons. This addressed the first barrier to NLP adoption – the need for involvement of an expert computer scientist. In order to address the second barrier – the time it takes to develop NLP tools – we propose to leverage high portability of NLP tools developed for the Canary platform, which can be easily reused on any other Canary installation. Below we describe the Canary NLP Tool Library that creates a forum for researchers to publicly share their NLP tools to allow everyone to build on accomplishments of others.

Software Description
The Canary software allows users to model their target information using lexicons and detection criteria, ranging from simple to complex, via a user-friendly graphical interface. The components forming the NLP pipeline include (1) text normalization & preprocessing; (2) vocabulary definition; (3) information discovery parameter definition; (4) structured value extraction; and (5) data filtering. Unicode support allows it to work with most languages.

To assist with data acquisition, a companion tool called the Canary Data Converter allows the conversion of data from various sources into a Canary-compatible format, with the possibility of adding support for new formats. After processing the input data, Canary produces structured output which can be analyzed to obtain study results ranging from prevalence statistics to longitudinal analysis of patient outcomes. Canary is now available for download at http://canary.bwh.harvard.edu.

Canary NLP Tool Library
Canary website includes a Library section where publicly shared NLP tools are posted. Information about each tool can include the concept(s) that the tool identifies in narrative data, its authors, accuracy and relevant publications. NLP tools can be submitted to the Canary website administrator for screening and posting. Security risks are minimal as an individual Canary NLP tools is a set of text-based configuration files describing the language model for identification of the concept of interest.

At this time Canary NLP Tool Library includes five NLP tools: a) Diagnosis; b) Family History; c) Left Ventricular Ejection Fraction; d) Aortic Valve Area and e) Mean Aortic Gradient. While Diagnosis and Family History tools are built for specific examples (diagnosis of calciphylaxis and family history of breast cancer, respectively), they can be easily adapted to other use cases. Recall and precision of the posted tools are estimated to range from 70-80% (Diagnosis and Family History) to as high as 95-98% (Left Ventricular Ejection Fraction and Aortic Valve Area). While it is likely that the shared tools may require modifications when transferred to another institution, these are expected to be minor, compared to the overall workload of NLP tool development. The goal of the NLP Tool Library is therefore to enable researchers to see farther ahead by standing on the shoulders of the giants – their colleagues.
Text De-Identification Impact on Subsequent Machine Learning Applications

Gary Underwood, MS¹, Andrew Trice, BS¹, Youngjun Kim, PhD², Jean-Karlo Accetta, MS¹, Stephane Meystre, MD, PhD¹,²
¹ Clinacuity, Inc., Charleston, SC; ² Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC

Introduction: The recent adoption of Electronic Health Record systems in the U.S. has resulted in very large quantities of clinical data becoming available in electronic format, with tremendous potential, but also equally growing concern for patient confidentiality breaches. De-identification of clinical data has been proposed as a solution to both facilitate secondary uses of clinical data and protect patient data confidentiality. The majority of clinical data is recorded as text, and efficient de-identification of this data can be based on Natural Language Processing (NLP) methods. Recent advances in NLP methods have been enabled by supervised machine learning and deep learning approaches, both relying on annotated clinical text for development and evaluation, on large quantities of clinical text. To allow use and sharing of this text, collections of clinical text have been de-identified according to the HIPAA Safe Harbor method, detecting all identifiers and replacing them with tags (e.g., “J. Smith” with [**PatientName**]) or with realistic surrogates (e.g., “J. Smith” with “A. Jones”), a process we call “identifier resynthesis.”¹ For NLP efforts focused on automatic text de-identification, the identifier surrogates should be as realistic as possible, maintaining the original morphology (capitalization, punctuation, spacing, etc.), variety, and consistency (e.g., all mentions of a patient name replaced with another same random name). Identifier resynthesis is the only viable option, but how this process possibly affects the accuracy of algorithms trained and tested with such de-identified text has rarely been evaluated, and only for sequence modeling algorithms such as Conditional Random Fields.²

Methods: To assess the potential impact of text de-identification with identifier resynthesis on subsequent use of the de-identified text for NLP-based named entity recognition (NER) using machine learning algorithms, we used an existing corpus of 228 clinical text notes from the Medical University of South Carolina. Notes were of the five most frequent types (e.g., progress note, discharge summary). All notes had been manually annotated for all identifiers defined in the HIPAA Safe Harbor rule³ and the corpus was split in a training subset (178 notes) and a testing subset (50 notes, based on power analysis). For resynthesis of the manually annotated identifiers, we used a component adapted from the BoB text de-identification application,¹ itself adapted from MIST.⁴ Identifier resynthesis breadth, accuracy and consistency were improved. The NER application we tested focused on identifiers and included part of CliniDeID, a complete text de-identification system. It included supervised machine learning algorithms (Conditional random fields, Margin infused relaxed algorithm and Structural support vector machines) but not the deep neural network components. We trained and tested it with various combinations of the original clinical notes (i.e., not de-identified) and the de-identified notes: training and testing with subsets both original or de-identified, or training with one and testing with the other. For accuracy measurement, metrics included recall, precision and the \( F_1 \)-score.

Results: All possible combinations of training or testing with original (“Original”) or de-identified text (“Resynth.”) were compared, and no significant difference was observed (Chi-squared test with \( p \) values between 0.998 and 0.912).

<table>
<thead>
<tr>
<th>Table 1. Accuracy evaluation results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train/Test corpora</strong></td>
</tr>
<tr>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Original-Original</td>
</tr>
<tr>
<td>Resynth.-Resynth.</td>
</tr>
<tr>
<td>Original-Resynth.</td>
</tr>
<tr>
<td>Resynth.-Original</td>
</tr>
</tbody>
</table>

Acknowledgements
Research supported by the U.S. National Institute for General Medical Sciences (R41GM116479)

References
**Title:** One Year After the Big Bang: “Things are going ok”  
**Authors:** Kim M. Unertl, PhD, Joyce Harris, MA, Shilo Anders, PhD, Peggy Cunningham, Laurie Novak, PhD, Taylor Avery, Carrie Reale, Patricia Sengstack, DNP, Nancy Lorenzi, PhD

**Affiliation:** Vanderbilt University Medical Center, Nashville, TN

**Introduction.** Health information technology (IT) implementations may “go live” on a single day, but the transition to effective use and adoption occurs over an extended time as individuals and their organization adapt. After 30+ months of planning, preparation, and training, Vanderbilt University Medical Center applied a “big bang” approach to implement a new health IT infrastructure, called eStar, across the entire clinical enterprise on November 2, 2017, involving 3 hospitals, 800 outpatient clinics, 125 physical locations, and 18,000+ users. Our research team studied the impact of the new Epic-based system on workflow, roles, and communication before, during, and after implementation. This poster reports on the final phase of our yearlong study of the health IT implementation.

**Methods.** Our team applied a systematic, rigorous approach to studying the large-scale health IT implementation. We developed a survey instrument for data collection one-year post-implementation. The research team collaboratively developed potential questions related to workflow and system use. The instrument was pilot-tested with a convenience sample, revised, and used during observation to assist with conversations about the new health IT and its ongoing impact on work and teams. The final survey instrument consisted of four Likert-scale questions (Table 1) and a free-text response question: “One year after eStar implementation started, how do you feel about how things are going?”

**Results.** We collected 171 surveys across 12 clinical areas (inpatient n=57, outpatient n=67, emergency department n=47). Respondents included patient services staff, medical assistants, nurses, advanced practice nurses, clinical leaders, physicians, and other staff. Of our sample, 44 people had prior Epic experience, 122 people did not have prior experience with Epic, and 5 people chose not to answer. Table 1 shows the results for the first four questions.

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Agree n (%)</th>
<th>Agree n (%)</th>
<th>Neither Agree nor Disagree n (%)</th>
<th>Disagree n (%)</th>
<th>Strongly Disagree n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am able to efficiently completely tasks using eStar</td>
<td>30 (17.5)</td>
<td>101 (59.1)</td>
<td>27 (15.8)</td>
<td>11 (6.4)</td>
<td>2 (1.2)</td>
</tr>
<tr>
<td>The tasks that I need to do in my work are easy to do using eStar</td>
<td>23 (13.5)</td>
<td>91 (53.2)</td>
<td>33 (19.3)</td>
<td>21 (12.3)</td>
<td>3 (1.8)</td>
</tr>
<tr>
<td>I am comfortable using eStar in my day-to-day tasks</td>
<td>29 (17.0)</td>
<td>121 (70.8)</td>
<td>13 (7.6)</td>
<td>7 (4.1)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td>eStar helps our team deliver high quality patient care</td>
<td>22 (12.9)</td>
<td>73 (42.7)</td>
<td>53 (31.0)</td>
<td>19 (11.1)</td>
<td>4 (2.3)</td>
</tr>
</tbody>
</table>

Answers to the free-text question (n=120) provided positive and negative feedback on individual and team experiences. Positive comments included an inpatient nurse who wrote, “It has been great. It is easy to use, complete tasks, and document.” Negative comments included an inpatient nurse who wrote “There are still a lot of kinks to be worked out. It’s not entirely user-friendly and physicians aren’t able to use it either, which makes our job more difficult.” Continued optimization needs, along with frustration with the optimization process were major themes. One inpatient physician wrote, “Needed improvements are not being implemented.” while an outpatient nurse noted, “The improvement process is tedious and bureaucratic.” Nurses and front desk staff also noted that frequent changes, sometimes without notice, made it difficult to become comfortable using the system. Across all roles and settings, people discussed usability issues, including the clicks required to complete tasks compared to the previous system, difficulty finding information, redundant/duplicated charting, difficulty charting common activities such as education, and scheduling inefficiencies. Another theme was an ongoing need for more training and help with personalization, as one outpatient nurse wrote, “There are so many things that Epic can do but we have had little to no training since initial implementation.” Finally, respondents discussed their expectations moving forward, as one physician wrote, “Seems to be ‘status quo’ at this point. Need to get people back to innovation/improvement mode.”

**Discussion and Conclusion** One-year post-implementation, adaptation to the new health IT infrastructure is still ongoing. While the majority of respondents agreed that they can complete day-to-day tasks efficiently with the new system, there was less agreement that the new system helped deliver high quality patient care. Analysis of free-text responses identified potential areas to focus attention and interventions on moving forward.
Refactoring and Expanding the Informed Consent Ontology (ICO)

Jonathan M. Vajda, MA1, J. Neil Otte, PhD2, Cooper Stansbury, MS3,
Marcelline R. Harris, PhD3, Frank J. Manion, PhD3, Cui Tao, PhD4
1University at Buffalo (SUNY), Buffalo, NY; 2Johns Hopkins University Applied Physics Laboratory, Laurel, MD; 3University of Michigan, Ann Arbor, MI; 4University of Texas Health Science Center at Houston, Houston, TX

The Informed Consent Ontology (ICO) represents the domain of informed consent, including the process of informed consent and related processes (e.g. informing a patient of their rights), the information produced by consent, and the resulting consent powers. Since sensitive materials such as specimens and data change hands across organizations governed by different policies and regulations, tracking the moving parts is critical to consent provenance. Our work addresses the need for ICO to have greater generality and granularity for widespread use as a reference ontology.

To improve this representation, the developers coordinated with the Document Acts Ontology (D-Acts), Ontology for Biobanking (OBIB), the Common Rule Ontology (CRO), and the Data Use Ontology (DUO). D-Acts represents document acts (e.g., signing contracts) and deontic roles (e.g., a role permitting a physician to perform surgery). The Ontology for Biobanking (OBIB) extends the Ontology for Biomedical Investigations (OBI), focusing on annotation terms relevant to biorepositories. CRO represents the Common Rule, DUO represents consent codes to annotate data with usage permissions. We used common refactoring methods applied in software development whereby code is improved through restructuring while preserving functionality while it is expanded and reevaluated. We reviewed every ICO term, marking for revision or deprecation in a spreadsheet. Use cases tested design flaws and maturity. Coordination with other ontologies promoted interoperability and avoided redundancy. We annotated blank consent forms with classes to show coverage of relevant terms.

ICO’s new release is improved for use as a reference ontology. The most general terms were revised in tandem with ontologies being importing in whole (denoted by black arrow in Figure 1) or in part (yellow arrow). Entities too specific for a reference ontology were recast or deprecated. The model of informed consent forms and their parts, e.g., directives, is enhanced. ICO incorporates CRO (purple arrow), enabling better representation of types of regulations across contexts. ICO’s new term ‘stasis of regulation’ is attractive to biobanks interested in tracking inventory over time or in querying when and where regulations are in force. This poster shows hierarchies of ICO and collaborating ontologies, and samples for intended use.

Acknowledgements: Support in-part by the National Human Genome Research Institute of the National Institutes of Health under award number 5U01HG99454 (Tao, PI), and a Michigan Institute for Data Science’s (MIDAS) Challenge Award (Harris, co-PI). The content does not necessarily express the official views of the National Institutes of Health or the University of Michigan.

References

6. https://github.com/EBISPOT/DUO
Mind the Gap: Revolutionizing the EHR Downtime Experience with an interoperable Workflow Tool

Patrick van Nieuwenhuizen, MD1, Jacqueline Soegaard Ballester, MD2, C. Priya Joshi, MD2, Subha L. Airan-Javia, MD3

1Hospital of the University of Pennsylvania, Philadelphia, PA  2VAMC, Philadelphia, PA, 3Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA,

Background

During EHR transitions or updates, hospital systems often undergo 8-16 hour EHR downtimes during which there is reduced access to data, documentation, and task management systems1-2. These downtimes have the potential to interrupt patient safety. At hospital systems that we are aware of, there are no alternative electronic platforms to fill the gap during downtimes.

Purpose

We created a web-based tool used for rounding and team-based task management that auto-populates data (labs, vitals, imaging reports), allows multiple users to manage interactive to-do lists and notes simultaneously, and is accessible via computer and phone. Though this tool was not initially intended to address the issue of data availability during downtime, we reviewed data usage during downtime to determine whether workspace tools have the potential to patch the discontinuity that occurred during the downtime.

Description

Six months after we implemented our web-based tool, our institution implemented a new EHR, during which there were a total of 27 hours of downtime across 4 hospitals spread across two days. During this downtime, the old EHR was viewable but did not have updated vitals, imaging, or lab data. In contrast, our tool, in addition to the static data at the time the old EHR became read-only, our tool continued to pull a subset of updated data from feeds directly from source systems including labs, radiology, clinical care plans, and tasks. In addition, while the old EHR was in a “read-only” mode during the downtime, clinicians were able to use our electronic tool to take notes and update tasks. After the new EHR went live, users were able to import this written care planning information into the new EHR. After the EHR transition, we assessed whether there was increased demand for our tool during the downtime. To do this, we measured the total number of different actions undertaken by users within our tool before, during, and after the downtime. These actions included those related to task management (figure 1) and viewing data (figure 2). Figures 1 and 2 illustrate that during the downtime, users accessed our application to view clinical data and manage their tasks more frequently than before or after the downtime. This increased usage suggests that there was demand for an alternative platform during the EHR downtime. Qualitatively, in user interviews, users reported decreased anxiety about the downtime because they were able to access clinical data within our tool. This may have improved patient safety during the transition, although we do not have data to validate that claim.

Conclusions

It is critical to think of how we can preserve safe patient care during EHR downtimes. Our experience suggests a need for alternative platforms for viewing data and managing tasks during EHR downtimes. Best practice may include the availability of alternative platforms for viewing data and managing workflow to stand in while EHRs are unavailable during downtimes.

References

OMOP Based Data Pipelines for Consistent Results in Cohort Identification Tools

Mahesh Vangala, MS; Yurima Guilarte-Walker, MS; Douglas Buell, BS; Keith Pelletier, BS; Katherine Luzuriaga, MD; Jomol Mathew, PhD.
University of Massachusetts Medical School, Worcester, MA.

Introduction and Background: Electronic Health Records (EHR) are increasingly becoming an integral part of healthcare. Cohort identification, feasibility studies and retrospective insights leveraging EHR data are accessible to researchers via nationally funded open-source solutions such as i2b2 and ACT as well as commercial third-party systems such as TriNetX. However, as a result of the wide variety of EHR systems available in the market and due to the proprietary nature of EHR data mapping, each organization is left with implementing custom Extract-Transform-Load (ETL) processes for loading data into these tools. This type of data loading is resource intensive and is especially cumbersome for healthcare organizations who employ EHRs from multiple vendors. A loading strategy that leverages a centralized and standard data model offers an optimal solution to enable easy data loading and establish consistent results across the systems.

Methods: We developed a dataflow pipeline (Figure 1) based on OMOP – CDM² (Observational Medical Outcomes Partnership – Common Data Model). OMOP was selected over other data models such as PCORnet as it is widely supported by cohort identification tools (TriNetX, i2b2 etc.). Nonetheless, we faced challenges integrating ACT/i2b2 ontologies with OMOP. ETL tools (Pentaho v6.0.0) and SQL scripts were used to map clinical data from the EHR source system (EPIC Clarity 2018) following OMOP – CDM specifications (v5.3). Prior to loading the data into each specific cohort identification tool’s data repository, a de-identification process was performed as per HIPAA guidelines (Section 164.514(a) Deidentification Standard - Safe Harbor Method Sections 164.514(b)(2)). Data mappings were evaluated using common metric counts (such as counts for Diagnoses, Procedures, Medications) between OMOP and the source EHR system – EPIC Clarity. In addition, we performed QA on randomized concept counts across our central research data warehouse, OMOP, TriNetX and i2b2 to confirm the data consistency. The resulting de-identified OMOP CDM Data repository was then used as the common data source for the cohort identification tools – i2b2 (v1.7.09c), ACT-SHRINE (v1.25.4) and TriNetX.

Results: We have successfully mapped the proprietary concepts from the EHR system and translated them into OMOP – CDM standardized concepts, namely, demographics and conditions (SNOMED), medications (RxNorm), labs (LOINC codes) and procedures (CPT4 codes). We then integrated data for a time period of October, 2017 to March, 2019 from 3,024,893 patients from our research data warehouse into the systems. Due to the proprietary nature of EHR system (EPIC – Clarity DB Schema), scripts cannot be shared over public servers. It is for this reason, we would like to share the intellectual material (non-proprietary scripts and/or guidance) in a controlled manner from UMCCTS Data Science Core.

References:

Acknowledgement: Research reported in this poster was supported by NCATS of the National Institutes of Health under award number: UL1TR000161 (KL, JM). The content of this publication does not necessarily represent the official views of the National Institutes of Health.
Patient Perceptions of Data Visualizations in Diabetes Self-Management

Jacqueline Vaughn BSN RN1; Allison A. Lewinski PhD, MPH, RN2; Anna Diane BSN RN1; Janee Stevenson BSN, RN3; Ryan J. Shaw PhD RN1

1Duke University School of Nursing, Durham, NC; 2Center of Innovation to Accelerate Discovery and Practice Transformation, Durham VA Health Care System, Durham, NC; 3Winston-Salem State University

Introduction: Mobile Health (mHealth) technologies such as wearables and smart phone applications (apps) help adults with type 2 diabetes engage in consistent, complex diabetes self-management. However, these technologies collect a significant amount of patient-generated health data (e.g., blood glucose, weight, physical activity). This large amount of data can cause data overload and/or fatigue, and loss of important information for both patients and clinicians resulting in poorer health outcomes (Lor, Koleck, & Bakken, 2019). Developing effective, meaningful, and interpretable ways to visualize these data addresses these challenges by providing valuable insight to guide patient decision making and self-management behaviors.

Methods: We conducted a mixed methods exploratory study for 60 patients with type 2 diabetes. Patients used 4 mHealth devices (i.e., wireless glucometer, cellular scale, wearable accelerometer, and smartphone delivered text message surveys) for 6 months to monitor and track their blood glucose, weight, physical activity, and medication adherence. We purposively selected 20 participants based on study engagement to participate in semi-structured interviews. As part of the study the research team developed data visualizations (e.g., charts and graphs) of these data for this subset of 20 study patients (Figure 1). The visualizations were shared with these patients retrospectively after they finished the study. The interviews focused on the patient’s participation in the study and the data visualizations.

Results: We identified two main themes related to the patients’ perceptions of the data visualizations. Patient responses about the usefulness of the visualizations were varied. Most patients (n=12), when looking at the visualization, could verbalize the relationship between the data points and their health actions. The first theme, meaningfulness of the data visualization, described how patients made sense of their healthcare data in relation to their self-management behaviors. We noted two sub-themes: (1) good to see the data—which stated that seeing the data was valuable because the patients could see the trends over time and obtained insight into their health behaviors; and (2) discrepancies—which the patient felt the visualization did not accurately depict their actions. The second theme, construction of the data visualization, described patient perceptions of the design and structure of the data visualization. Patients stated three types of suggestions for the data visualizations: requests for additional information in the visualization, various layout preferences, and access issues.

Conclusion: Effective visualizations can illustrate comparisons, associations, and trends over time in a succinct and easily understood manner. These results are the first step in understanding the differing needs of visualizing data for patients with diabetes and their healthcare providers. Our findings indicate the importance of incorporating patient feedback in the design of the visualization in order to ensure this feedback is actionable and assists the patient.

Time Expression Annotation for Multiple Clinical Specialties

Natalia Viani, PhD¹, Hegler Tissot, PhD², Ariane Bernardino, MS¹, Sumithra Velupillai, PhD¹

¹Institute of Psychiatry, Psychology and Neuroscience, Kings College London, London, UK; ²Institute of Health Informatics, University College London, London, UK

Introduction

When applying natural language processing (NLP) methods to the analysis of clinical notes, temporal information extraction is a crucial yet challenging task. The availability of temporally annotated corpora is an important step towards the development of well-performing and generalizable NLP tools. In the clinical domain, two shared corpora have been created including temporal annotations: the 2012 i2b2 corpus¹ and THYME². In this work, we present a corpus of clinical texts annotated with time expressions. This corpus will be made available to the NLP community.

Methods

We extracted documents from MTSamples (www.mtsamples.com), a collection of Medical Transcription Sample reports for multiple specialties. We selected the following specialties for manual annotation: discharge summaries, psychiatry-psychology, pediatrics, and emergency. These subsets were double-annotated for five types of time expressions (date, time, duration, frequency, and age-related), including their normalized values. As a proof-of-concept for NLP development, we applied SUTime (a rule-based system for time expression extraction) on annotated documents³.

Results

For each clinical specialty, Table 1 reports the number of documents (with total number of tokens), the number of time expressions marked by at least one annotator (merged), and those marked by both annotators (overlap). For the inter-annotator agreement (IAA), we report F1 score for text spans (allowing overlapping annotations) and type accuracy (on overlap annotations). In all document subsets, the most frequent “type disagreement” was duration-date. For SUTime, we report precision (P), recall (R), and F1 score on merged annotations (without system refinement).

Table 1: Time expression annotation and SUTime results per clinical specialty.

<table>
<thead>
<tr>
<th></th>
<th>dis. summaries</th>
<th>psych.</th>
<th>pediatrics</th>
<th>emergency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>108 (55K tokens)</td>
<td>53 (68K tokens)</td>
<td>70 (37K tokens)</td>
<td>75 (52K tokens)</td>
</tr>
<tr>
<td>Time expressions (merged - overlap)</td>
<td>1,378 - 994</td>
<td>1,227 - 840</td>
<td>566 - 360</td>
<td>801 - 496</td>
</tr>
<tr>
<td>IAA F1 (type acc.)</td>
<td>0.84 (0.91)</td>
<td>0.81 (0.88)</td>
<td>0.73 (0.71, 0.72)</td>
<td>0.76 (0.86)</td>
</tr>
<tr>
<td>SUTime P, R, F1</td>
<td>0.85, 0.73, 0.78</td>
<td>0.73, 0.71, 0.72</td>
<td>0.73, 0.84, 0.78</td>
<td>0.79, 0.74, 0.76</td>
</tr>
</tbody>
</table>

Conclusion

We present a corpus of documents annotated for time expressions, considering four clinical specialties. We will use the annotated corpus to: 1) study how temporal information is documented in clinical text, with a focus on disagreements involving durations, and 2) define new ways to capture the most useful information for clinical timeline reconstruction. The SUTime results show that our corpus can be successfully exploited for NLP system development and evaluation.

References

Enabling Data Democratization - Interactive Graphical Representations of Unplanned Readmissions for Cancer Patients

Cyd Charisse Villalba, MS, RN-BC, PMP, Sara Berger, MBA, Steve Martin, MD, Peter Stetson, MD, MA, Wazim Narain, PhD, MPH, Joery J. Nunez, BA, Tiffanny Newman, PhD, Patrice Schwegman, BS, Rita Reynolds Silverberg, MS, Patrick Samedy, MS, MBA, and Dhruvkumar Patel, MS
Memorial Sloan Kettering Cancer Center, New York, NY

Introduction
Cancer is the second leading cause of death in the United States, with an incidence rate increasing yearly due to an aging population. The Agency for Healthcare Research and Quality (AHRQ) approximates $80.2 billion was spent directly on cancer-related medical costs in 2015, 38% of which was incurred during inpatient hospitalizations. Current non-cancer readmissions metrics focus on 30-day all-cause, disease-specific (AMI, COPD, HF), or post-procedure (vascular, PI) readmissions. The Comprehensive Cancer Center Consortium for Quality Improvement (C4QI) along with the Alliance of Dedicated Cancer Centers (ADCC), which includes Memorial Sloan Kettering Cancer Center (MSKCC), collaborated with the National Quality Forum (NQF) to develop NQF 3188: 30-Day Unplanned Readmissions for Cancer Patients. Approved in 2019’s final ruling, it is now a required claims-based outcome measure for FY 2021. Notwithstanding regulatory compliance, the establishment of a cancer-specific readmissions metric enables a nuanced assessment of the specific factors and interventions that are most meaningful to this cohort. Medicare cancer readmissions are higher than all-cause readmissions within non-cancer institutions, and even within cancer centers, specific surgical treatments and cancers, as with bladder, pancreatic, ovarian and liver cancers, the risk of readmissions are higher (likely due to post-surgical complications, infection and dehydration). Other significant predictors of cancer readmissions include comorbidities, older age, advanced disease, and index length of hospital stay. In light of the hallmark regulation and aligning with MSKCC’s mission, the team established an unplanned cancer readmissions data source enabling intuitive and robust operational and analytical dashboards.

Methods
The in-house development team augmented a pre-existing inpatient readmissions datamart with discharge status, disposition, age, primary payor, and diagnosis, to enable application of measure criteria. Readmission rate was then trended monthly from 2017 to the most recent month with a complete 30-day post-discharge period. In addition to the 30-day metric, 7 and 14-day readmission rates were also displayed to assess whether these are more sensitive readmission markers. Other dimensions such as discharging department, service, attending and unit were also graphically presented to enable drill-down analysis. Moreover, a separate tab outlining comprehensive case details allowed for ad-hoc patient-level analysis. Lastly, the readmissions metric was integrated into high-level operational and performance dashboards, ensuring key stakeholders are aware of trends and patterns.

Figure 1. Cancer-patient readmission rate by primary payor over time.

Results
In addition to providing a baseline in preparation for regulatory compliance, the dashboard highlighted marked variances in readmissions rate and patient volume between discharging departments, services, attendings and units. These findings may aid in outlining the most appropriate benchmarking methods to apply inter- and intra-institution. Also, enabling trended comparisons between Medicare and non-Medicare cohorts for three timed metrics (7, 14, and 30-day) may be useful in providing much-needed insights into the cancer patient population at MSKCC. Further analysis is required to establish correlation, causation and clinical significance between different factors that could impact readmissions.

Conclusion
Aligning a regulatory project deliverable to a broader institutional initiative allowed one data product to fulfill multiple needs, current and future. Next steps include expanding the underlying data source to allow for broader and deeper analyses, as well as tailoring subsequent design and content for specific workgroups and use-cases.

References
Dynamic Phenotyping to facilitate Accrual for Prospective Clinical studies: A Case Study in Heart Failure

Kavishwar B. Wagholikar MBBS PhD1,2,3, Christina M. Fischer MS2, Alyssa P. Goodson MS4, Christopher D. Herrick MBA4, Taylor E. Maclean2, Katelyn V. Smith2, Liliana Fera MSN2, Thomas A. Gaziano2, Jacqueline R. Dunning2, Joshua Bosque-Hamilton2, Lina Matta2, Eloy Toscano, MS,4 Brent Richter MBA,5 Layne Ainsworth MBA, 4 Michael F. Oates4, Samuel Aronson ALM, MA4, Calum A. MacRae MD PhD1,2, Benjamin M. Scirica MD1,2, Akshay S. Desai MPH1,2, Shawn N. Murphy MD, PhD1,3

1Harvard Medical School, Boston, MA; 2Brigham and Women’s Hospital, Boston, MA; 3Massachusetts General Hospital, Boston, MA; 4 Partners Healthcare, Boston, MA

Introduction

The adoption of electronic health records (EHRs) in the United States has generated large datasets that enable observational research. An essential step in utilizing these datasets for research is to identify patients with certain characteristics that match a set of eligibility criteria—a process known as phenotyping or patient cohort identification. We describe a dynamic phenotyping approach to facilitate recruitment for a prospective study on heart failure.

Methods

Dynamic phenotyping consists of multiple computer-assisted screen cycles. At the start of each cycle, the phenotyping algorithm is used to identify eligible patients from the EHR, creating an ordered list such in which those patients most likely to be eligible are listed first. This list is manually screened, and the results are analyzed to improve the phenotyping algorithm for the next cycle. We applied this approach to EHR data at Brigham and Women’s Hospital (BWH) to accrue heart failure patients for an intervention study.

The eligibility criteria for the study included adult patients that had a diagnose of heart failure with reduced ejection fraction, and whose cardiac care was managed primarily by a cardiologist at BWH. We utilized a variety of algorithms for identifying the patients including rules, regular expressions for text processing, and logistic regression on coded data and concepts extracted from clinical notes. For evaluation, in each cycle we measuring the positive screen rate (PSR) using the annotations from the manual screening in the next cycle (Table 1).

Results and Discussion

A total of 1,022 patients were screened in four cycles, with 223 (23%) of patients being found eligible for enrollment into the intervention program. When a single run of computer pre-screening is used, the pool of eligible patients is expected to decline as manual screen progresses, resulting in more false positives and a subsequently downward trend in the positive screen rate (PSR). We used multiple computer pre-screens and improved the computer algorithm for each pre-screen. This process offset the dip in the PSR by providing an enriched cohort in every cycle. Without an iterative approach, the PSR was expected to dip below the 20% measured in the first cycle; however, the cyclical approach increased the cumulative PSR to 23%. Our results demonstrates that dynamic phenotyping can improve recruitment for prospective clinical studies. However, additional case studies are required to establish our finding.

References


Acknowledgements

This study was supported by NIH grants R01-HG009174 and R00-LM011575, and funding from Partner Healthcare.
Integrative workflow for lung cancer clinical decision support: needs and opportunities

Jinglu Wang, MSc, Yalini Senathirajah, PhD
1Weill Cornell Graduate School of Medical Sciences, New York, NY, USA; 2University of Pittsburgh, Pittsburgh, PA, USA

Introduction
Incorporation of clinical decision support (CDS) at the point-of-care requires integration of multiple information sources, involvement of system vendors, consideration of population data on which models may be based, and interoperability, as information may be sent outside a health system or vendor product. Currently, there are several lung cancer diagnostic and risk prediction tools online. All require the provider to perform the cognitively loading and time-consuming task of data gathering (from EMR, patient, other providers, external systems, and paper records) and insertion into (usually) online forms-based systems which then return a set of risk predictions and further recommendations. Acting on these usually involves further manual documentation, explanation to patient and other providers, order entry, and communication to other stakeholders.

Interoperability and current technical approaches
As a majority of tools are realized as online forms, the public APIs that are now required in EMRs can be leveraged to gather and send data simultaneously to multiple tools (e.g. Fleischner, Lung-RADS, NCCN) and receive back their recommendations in a common interface, allowing comparison and order entry on the provider-facing side. FHIR as both a data model and a communication standard would permit appropriate visualization and explanatory interfaces such that providers with little or no knowledge of the specific algorithms can progressively view rapid, at-a-glance status, or deeper, more visible explanation, as desired. This ‘white box’ approach would help ensure the ‘five rights’ CDS requirements (Right Information, Person, Intervention format, Channel, and Time in workflow) are met. This approach permits development of models with better AUC (accuracy of 73% for traditional clinical diagnosis vs. 85% with cognitive computing approaches). It would make use of CDS hooks, specialized patient-directed questionnaires, and NLP for other information forms such as faxes and transmitted notes. As the science changes, health information technology needs to leverage a variety of information in order to implement effective LDCT screening at the population level. There is still tremendous variation in the use of guidelines in clinical settings. Radiologist concordance on imaging studies alone vary as much as 62% on part-solid nodules to 85% on solid nodules. While clinical risk calculators are helpful, they are seldom used in the evaluation of patients due to the need for manual data entry. Recent randomized trials such as the PLCO and NLST have shown that current guidelines are suboptimal algorithms for all but the same underlying populations. Several different algorithms exist: 1. NPS-BIMC (Bayesian Inference Malignancy Calculator); 2. Brock University Calculator 3. Solitary Pulmonary Nodule Malignancy Risk (Mayo Clinic model) 4. MSKCC Lung Cancer Screening Decision Tool 5. AHRQ Lung Cancer Screening Tools for Clinicians and Patients

Conclusion
Predictive analytics holds promise but needs to facilitate patient-provider-team interaction. Newer interoperable technology combined with flexible design may allow us to align care and technology to create provider-centered workflows; only when this is done is widespread use likely and achieve the promise of precision oncology.

References
CLAMP-ATR: A deep learning pipeline for attribute recognition of clinical concepts

Jingqi Wang, MS, Yaoyun Zhang, PhD, Hua Xu, PhD,
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, United States

Introduction: Attributes of clinical concepts, such as severity and negation modifiers of clinical problems in clinical text, are vital elements for clinical natural language processing (NLP). A high-performance, easy-to-use toolkit to automatically recognize attribute information of clinical concepts from clinical text is highly desirable. Here we introduce CLAMP-ATR (Attribute), a newly developed comprehensive pipeline for attribute recognition with the CLAMP NLP toolkit.¹ It not only achieved high-performance on attribute recognition across different corpora, but also made customization easier by providing a Docker container for training deep learning models and a GUI interface for building pipelines.

Methods: Currently, CLAMP-ATR recognizes three types of important clinical concepts and their attributes: 1) Problem and modifiers of subject, negation, severity, body location, etc.; 2) Lab test and its value; 3) Medication and signatures of dose, form, rout, frequency, etc. An NER model is first built to recognize primary clinical concepts, including problems, tests, and medications. Then the CLAMP-ATR will use the output of primary concepts as features to recognize their attributes. To achieve the most optimized performance of the attributes, we developed three attribute models separately. Algorithm: We use the deep learning based biLSTM-CRF algorithm for both concept and attribute recognitions. Character embeddings and word embeddings are used as input features. The word embeddings are pretrained on the MIMIC III dataset. For attribute recognition, an additional embedding layer is added to represent the primary entities as features. Evaluation: The problem-modifier model was evaluated on the data of the SemEval 2015 task 14 challenge;² the test-value model was evaluated using 338 annotated discharge summaries from MTSamples; and the medication-signature model performance was evaluated on the N2C2 shared task track 2.

Model customization: We build a Docker container for training deep learning models using local data. It works as an insulated virtual environment with all deep learning libraries and pre-set parameters. The users can run this container on their own corpus with just simple docker commands. In addition, the trainer can use GPUs if they are available in the host machine. It also provides advanced options if the users need to update the pretrained embeddings or adjust model parameters. The final model will be encoded into a single file, which can then be loaded into a CLAMP pipeline, where users can further customize it using the CLAMP interface.

Results and conclusions: The performance on each type of clinical concepts and their attributes is listed in Table 1. Notably, the performances of problem-modifier and medication-signature recognition are comparable to the top ranked in the SemEval 2015² and N2C2 challenges. The performance of test-value recognition is also promising for practical applications. Figure 1 shows the visualization of identified concepts and their attributes in the CLAMP interface. A high-performance, easy-to-use toolkit to automatically recognize concept and attribute information from clinical text will greatly facilitate related clinical and translational research and applications.


Conflicts of Interest: Dr. Xu and The University of Texas Health Science Center at Houston have research-related financial interests in Melax Technologies, Inc

Table 1. Performance of concept and attribute recognition

<table>
<thead>
<tr>
<th>Types</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Types</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>problem</td>
<td>90.2</td>
<td>90.6</td>
<td>0.904</td>
<td>medication</td>
<td>93.7</td>
<td>95.3</td>
<td>0.945</td>
</tr>
<tr>
<td>negation</td>
<td>88.2</td>
<td>87.6</td>
<td>0.879</td>
<td>strength</td>
<td>94.9</td>
<td>94.6</td>
<td>0.948</td>
</tr>
<tr>
<td>severity</td>
<td>85.3</td>
<td>90.3</td>
<td>0.877</td>
<td>route</td>
<td>94.9</td>
<td>95.6</td>
<td>0.952</td>
</tr>
<tr>
<td>bodyloc</td>
<td>79.6</td>
<td>81.6</td>
<td>0.805</td>
<td>frequency</td>
<td>82.7</td>
<td>83.1</td>
<td>0.829</td>
</tr>
<tr>
<td>test</td>
<td>89.2</td>
<td>88.3</td>
<td>0.891</td>
<td>duration</td>
<td>68.9</td>
<td>72.1</td>
<td>0.705</td>
</tr>
<tr>
<td>value</td>
<td>97.7</td>
<td>95.5</td>
<td>0.966</td>
<td>form</td>
<td>95.5</td>
<td>93.5</td>
<td>0.945</td>
</tr>
</tbody>
</table>

Figure 1. Visualization in CLAMP
Association between Cardiotoxic Chemotherapy and Myocardial Infarction

Liwei Wang, MD, PhD1, Suzette J. Bielinski, PhD1, Paul A. Decker, M.S1, Jill M. Killian, B.S., Nicholas B. Larson, PhD, M.S1, Rui Zhang, PhD2, Hongfang Liu, PhD1
1Mayo Clinic, Rochester, MN, USA; 2University of Minnesota, Minneapolis, MN, USA

Introduction

Cancer treatment-related cardiotoxicity can occur due to direct effects on cardiac myocytes and may include damage to the vascular endothelium, hemodynamic flow alterations, and thrombotic events. Incidence of cardiotoxicity of cancer treatment vary from 7.4% to 13.3%. Assessment of the long-term effects (10–20 years) of chemotherapy on coronary heart disease is lacking. In a population based cohort, we sought to study the effect of exposure to cardiotoxic chemotherapy and subsequent coronary heart disease, defined as acute myocardial infarction (MI).

Methods

We assembled a cohort of 74,470 Olmsted county residents on 1/1/2006 with no prior history of MI using the Rochester Epidemiology Project 1 and followed them through 12/31/2016 for incident MI. To extract cardiotoxic chemotherapy exposure status prior to the index date, a four-step strategy was used. First, ingredients of all drugs these patients were exposed to before 1/1/2006 were extracted using a natural language processing tool, MedXN2. Ingredients were then identified as chemotherapy drugs by mapping to the chemotherapy ingredients provided by a knowledge base (https://hemonc.org/wiki/Main_Page). The ascertained chemotherapies were further linked to potential side effects using the SIDER knowledge base. Finally, the side effects were mapped to the System and Organ Class (SOC) of Medical Dictionary for Regulatory Activities (MedDRA) system and “cardiac disorders” SOC was used to extract cardiotoxic chemotherapies. The association of prior cardiotoxic chemotherapy exposure (ever/never) with risk of MI was assessed using proportional hazards regression, with corresponding hazard ratios (HR) and confidence intervals (CIs) reported.

Results

Of the 74,470 participants in the cohort, 6,587 (9%) were previously exposed to cardiotoxic chemotherapies. The subjects were 51±15 (range: 30-105) years old; 54% were female and 88% were white, reflecting the population of the upper Midwest. Average systolic blood pressure was 124±18 (57-252) mmHg, 22% received hypertension treatment in the previous 2 years, and 8% had a diagnosis of diabetes. Average total and HDL cholesterol was 195±38 and 56±17 mg/dL, respectively. There were 1494 (2%) incident MI during follow-up. There was a significant univariate association of exposure to cardiotoxic chemotherapy and subsequent MI. This association remained with further adjustment for age, sex, race, systolic blood pressure, hypertension treatment, and diabetes (Table 1). Similar results were observed when the analysis was stratified by sex.

Table 1. Results for the association of exposure to cardiotoxic chemotherapy and MI.

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>HR (95% CI, P)</td>
<td>N</td>
</tr>
<tr>
<td>Univariate</td>
<td>74,470</td>
<td>1.64 (1.41 - 1.90, &lt;0.001)</td>
<td>39,844</td>
</tr>
<tr>
<td>Multivariable</td>
<td>68,883</td>
<td>1.21 (1.04 - 1.40, 0.015)</td>
<td>37,271</td>
</tr>
</tbody>
</table>

Conclusion

Leveraging an existing cohort of 74,470 Olmsted county residents, we identified a significant association between exposure to cardiotoxic chemotherapy and subsequent MI. Future work will expand to investigate the association of chemotherapies and other cardiovascular diseases as well as explore mediating factors such as increased burden of cardiovascular risk factors in cancer survivors.

Reference

Design and Evaluation of a Decision Support Risk Calculator for Radiotherapy for Older Women with Early-stage Breast Cancer

Shi-Yi Wang, MD, PhD¹, Suzanne B. Evans, MD², Sarah Mougalian, MD², Brigid K. Killelea, MD¹, Liana Fraenkel, MD, MPH², Cary Gross, MD², Fuad Abujarad, MS, PhD²

¹Yale School of Public Health, New Haven, CT, USA; ²Yale School of Medicine, New Haven, CT, USA

Radiation therapy – or radiotherapy – is typically received by women with breast cancer after a lumpectomy to lower the risk of cancer recurrence in the affected breast.¹ The benefits of radiotherapy largely depend on tumor characteristics, comorbidities, and the patient’s functional status.² In fact, it has been found that radiotherapy can be safely omitted for a majority of older women aged 65 or older with early-stage breast cancer.³ In order to ensure that these women do not receive unnecessary therapy that could potentially cause undue harm, we developed a web-based mobile application for comprehensive risk calculation called “Radiotherapy for Older Women” (ROW). This tool provides individualized estimates of overall survival and local recurrence, accounting for patient age, comorbidity, functional status, and tumor characteristics. These estimates are presented in a readily interpretable manner, which patients and providers can use to discuss whether the patient should pursue or forgo radiotherapy.

ROW was developed using a User-Centered Design (UCD) approach. We first estimated individualized risks of breast cancer recurrence, breast cancer mortality, and all-cause mortality using simulation modeling. The simulation model integrated two existing prediction models, the Early Breast Cancer Trialist Collaboration Group prediction model for breast cancer specific outcomes and ePrognosis for life expectancy. We then convened an advisory committee comprised of breast cancer survivors, patient advocates, oncology clinicians, and researchers to receive feedback on the tool’s functionality. We also used expert interviews and stakeholders group meetings to develop a set of functional and technical patient-centered requirements. We incorporated users’ needs and limitations at each stage of the design process per the UCD approach and conducted User Experience (UX) evaluations as well to ultimately produce a usable and acceptable software that delivers a more satisfying user experience. To demonstrate feasibility, we have been testing the use of ROW with two groups of older women subjects: healthy volunteers and breast cancer patients. We have been gathering performance and satisfaction data to evaluate the extent to which the user interface function matches the user requirements in design specification, which will ensure that user needs have been met. It is possible that our tool could integrate with the EMR in the future. Currently, we are validating the tool without integration to ensure its feasibility.

To date, we have recruited 20 healthy volunteers and 5 breast cancer patients for the feasibility evaluation. Our initial analyses suggest that women are very satisfied with the tool and that 96% of women would recommend the use of this tool to others. The majority of women have stated that they liked the tool and that it was easy to use, and most thought that the functions in the tool were well integrated. A majority of women have also reported high confidence in their ability to navigate the tool on their own and have stated that it is easy to learn. At this stage, we feel it is important for the patient to use the tool under the supervision of the provider. These results suggest that ROW is feasible and satisfactory for older women with breast cancer, and there is potential for the tool to reduce barriers to decision support implementation and improve patient-provider communication. The study is currently ongoing to further validate this tool.

References
Trajectories of Functional Status of Cognitively-impaired Patients in EHRs

Yanshan Wang, PhD, Sunghwan Sohn, PhD
Division of Digital Health Sciences, Mayo Clinic, Rochester, MN

Introduction
The aging population has led to an increase in cognitive impairment (CI) resulting in significant costs to society. In general, CI is diagnosed by clinicians through assessment. However, it is not routinely performed, causing a delay in timely diagnosis. Currently, little is known about temporal trends of patient health functions (i.e., activity of daily living [ADL]) in electronic health records (EHRs) and how these trends are associated with the onset of CI in elderly patients. We investigated the use of EHRs to better understand early signals of elderly patient CI by examining ADL trajectories.

Materials and Methods
The study data consists of physician-diagnosed CI patients (n=1,435) and cognitively unimpaired (CU) patients (n=1,435) matched by age and sex from patients 65 years of age or older in the Mayo Clinic Biobank. The physician-diagnosed CI patients were determined based on diagnosis under the diagnosis section in clinical notes. The ADL was collected from the current visit information (questionnaires) and specific sections in clinical notes, such as instructions for continuing care, ongoing care orders, and system review. The clinical notes were processed by the MedTaggerIE (https://github.com/medtagger/MedTagger), a clinical NLP pipeline developed by Mayo Clinic, to extract ADL related concepts. We use the ADL categories as follows: basic ADL - bathing, dressing, transferring, toileting, and feeding; instrumental ADL - transportation, preparing food, housekeeping, and responsibility for own medications. The trajectories of ADL between CI and CU patients were compared in every 6 months for 5 years before the first physician-diagnosed CI and the latest visit for CI and CU patients, respectively. The ratio and trends of deteriorated ADL and/or patients were analyzed.

Results
CI patients had more deteriorated ADL than CU patients for all ADL categories. Fig 1 shows the ratio of deteriorated ADL for each category, 1 year before CI diagnosis for the CI group and the latest visit for the CU group. The most deteriorated ADL in 1 year prior was transferring (13% for CI and 10% for CU patients) in basic ADL and housekeeping (10% for CI and 7% for CU patients) in instrumental ADL. The instrumental ADL (a right half in Fig 1) was more deteriorated than basic ALD (a left half in Fig 1). The difference between the two groups is relatively small for feeding and transferring. The trajectory for responsibility for own medication was the steepest among ADL categories. Fig 2 depicts the deteriorated ratio of responsibility for own medication for the past 5 years (the dotted line denotes linear trend). As can be seen in Fig 2, CI patients have a much steeper slope when approach to the CI diagnosis than CU patients, which is notable much before the actual CI diagnosis.

Discussion
We have observed notable differences in trajectories of ADL mined from EHRs between CI and CU patients, approximately 1 to 1.5 year(s) earlier than actual physician-diagnosis CI. The trajectories of certain individual ADL (e.g., responsibility of own medication) were closely associated with the CI development and might have a potential to surrogate costly assessment, facilitating a routine practice (i.e., EHR data) in clinical care and research.
Clinical Use of an Information Retrieval Framework for Cohort Discovery from Electronic Health Records

Yanshan Wang, PhD1, Andrew Wen, MS1, Sijia Liu, PhD1, Jennifer St. Sauver, PhD1, Adil E. Bharucha, MD2,3, Chunhua Weng, PhD4, Hongfang Liu, PhD1
1Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; 2Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA; 3Center for Clinical and Translational Science, Mayo Clinic, Rochester, MN, USA; 4Department of Biomedical Informatics, Columbia University, New York, NY, USA

Background

The widespread adoption of Electronic Health Records (EHRs) has enabled secondary use of EHR data for clinical research and healthcare delivery. Many institutions have established clinical data warehouses (CDWs) in conjunction with cohort discovery tools, a majority of which are based on relational databases, such as i2b2, to enable investigators to use EHRs for cohort discovery in clinical research. As a significant portion of clinically relevant patient information is embedded in clinical narratives, cohort discovery using only structured data such as diagnosis codes or procedure codes will display limited retrieval performance. To compensate for this, substantial efforts have been made to use clinical narratives for cohort discovery; one major approach is the usage of natural language processing (NLP) to encode concepts from unstructured EHR data for cohort retrieval. However, existing NLP methods for concept encoding are either optimized for a single information extraction task or have imperfect generalized performance. This problem is magnified when the extracted NLP artifacts are utilized in downstream applications, such as concept matching for cohort criteria. Information retrieval (IR) is a field that is concerned with the structure, analysis, organization, storage, searching, and retrieval of relevant information. While IR has established its success in providing easy access to the vast amount of information available on the Web, IR has not been widely utilized in cohort discovery due to its lack of ability to handle complex EHR data and cohort criteria.

Method

A novel IR framework empowered by NLP is currently under-development at Mayo Clinic that will be able to store and index EHR data, and retrieve and rank patients for cohort discovery tasks using not only structured EHR data, but also unstructured clinical text. The IR framework is a toolkit independent of the data to be indexed, and thus it is generalizable and portable. Due to these innovations, the IR framework has an advantage over traditional cohort discovery tools for querying unstructured EHR data as well as an advantage over text-based search engines for querying both structured and unstructured EHR data given complex cohort criteria with contextual constraints. However, the challenge we have been facing are: 1) what would be the ideal clinical use cases of the IR framework, and 2) how to evaluate the framework.

This study aims to identify the clinical use cases for utilizing and evaluating our cohort discovery tool based on an information retrieval (IR) framework. Two practical clinical applications have been found: identifying patient cohorts for retrospective clinical studies, and recruiting patients for clinical trials. Retrospective cohort studies, one of the research areas that mostly rely on EHR data, have been used to evaluate population health and identify risk factors. Enabled by the Rochester Epidemiology Project (REP), many cohorts have been created with EHR data. Another clinical use case is clinical trial recruitment, which is fundamental to the success of randomized clinical trials since it remains the biggest barrier to clinical and translational research. Mayo Clinic ranks No. 1 among 51 academic institutions in terms of both number of clinical trials and number of results reported. The data of a number of completed clinical trials and recruited patients are available at Mayo Clinic to evaluate the approach.

Results

20 cohorts were identified from REP studies based on disease areas and study time. These cohorts covered different diseases, such as zoster, heart disease, and fracture. Another 20 completed clinical trials at Mayo Clinic were identified, for which the EHRs of these recruited patients were available at Mayo Clinic.

Conclusion

This study identified 40 cohorts from two clinical use cases, including identifying patient cohorts for clinical studies and recruiting patients for clinical trials, that could be potentially used to utilize and evaluate our IR framework.
Analysis and Resolution of Medication Terminology Errors to Achieve Semantic Interoperability between the Department of Veterans Affairs and Department of Defense

Pradnya P. Warnekar, MSHI1, Jaqui Parker, RPh1, Carol A. Bean, Ph.D1, Daniel Territo, MS1, Lauren C. Thompson, Ph.D1
1 VA-DoD Interagency Program Office, Rosslyn, VA

Introduction

The Clinical Data Repository/Health Data Repository (CHDR)1 is a health information exchange framework between the Department of Veterans Affairs (VA) and Department of Defense (DoD), used to transmit, in real-time, allergy and medication profiles for patients registered for care in both Departments by using national terminology standards. VA and DoD have their own processes to map local medication names to RxNorm2. The RxNorm name and RxNorm Concept Unique Identifier (RxCUI) are transmitted to the recipient Agency in a HL7 message. The receiving Agency translates the RxCUI back to an equivalent local code that participates in its existing clinical decision support systems. If the receiving Agency cannot translate the RxCUI, a terminology error is recorded, and the medication is not stored in the patient’s medical record. This results in exchange of an incomplete patient medication profile and physicians must spend time collecting and documenting the missing information. Most of the terminology errors are resolved "naturally" when the subsequent versions of mapping files are deployed in production, but interoperability suffers in the meantime and comparable errors recur in the new files. This paper analyzes the causes and rates of medication terminology errors in CHDR over the calendar year 2018 and suggests ways to resolve errors in a shorter time frame.

Materials and Methods

For the current analysis, CHDR production audit logs from January 01, 2018 to December 31, 2018 were queried for all medication terminology errors. All the RxCUIs that each Agency received and could not map back to one of its local codes were collected. A list of all unique RxCUIs with their frequencies was compiled for VA and DoD. Both lists were compared to the next versions of mapping files from both Agencies and RxNorm February 2019. The comparison allowed identification of errors that would be resolved by the next content updates at both Agencies. The errors that would not be resolved by the next content updates were divided into three groups: (1) Errors due to mapping rules differences (2) Erroneous mappings (3) Formulary differences between the two Agencies.

Results

In 2018, the total number of translations attempted at VA and DoD were 59,440,924 and 27,397,326 respectively. The number of failed translations at VA and DoD were 3,335,496 (1327 unique RxCUIs) and 1,159,114 (612 unique RxCUIs) respectively. When the latest mapping files are deployed in production, two types of content updates will occur: (1) retired RxCUIs will be replaced by active RxCUIs, and (2) the latest VA and DoD medication names, with their corresponding RxCUIs, will be available for exchange. These updates will resolve approximately 86% of DoD errors and 67% of VA errors. Of the unresolved errors, mapping rules differences and mapping errors account for 64% of VA and 78% of DoD errors. Differences between the two Agencies’ formularies cause 36% of VA and 22% of DoD errors. For example, VA carries the tablet form of Ibrutinib 140 MG while DoD carries the capsule.

Conclusion

To achieve and maintain a successful mediation terminology-based interoperability effort, common mapping rules, frequent, coordinated content updates, error monitoring and resolution, and open, regular communications are needed.

References

Scalable Process to Generate Aggregated Patient Data for Analysis

Nich Wattanasin, MS¹, Taowei David Wang, PhD¹, Vivian Gainer, MS¹
Shawn N. Murphy, MD, PhD¹,²
¹Partners Healthcare, Boston, MA; ²Massachusetts General Hospital, Boston, MA

Abstract
The Informatics for Integrating Biology and the Bedside (i2b2) platform is a framework enabling clinical research and data analysis at participating sites of the Accrual to Clinical Trials (ACT) network. At Partners Healthcare, we have developed a scalable and self-service approach for generating aggregated patient data to support the patient recruitment process for ACT.

Introduction
The goal of the National Clinical and Translation Science Award (NCATS) Accrual to Clinical Trials (ACT) project is to create a federated network made up of sites from the Clinical and Translation Science Award (CTSA) Consortium to accelerate participant accrual to the nation’s highest priority clinical trials. The network for cohort discovery utilizes the Shared Health Research Information Network (SHRINE) platform to connect CTSA hubs comprised of Information for Integrating Biology and the Bedside (i2b2) clinical data warehouses for research.

One of the aims of ACT is to enable the identification of potential research subjects for clinical trials among the ACT participants across the CTSA consortium. While initial cohort discovery and feasibility studies are made by obtaining aggregate patient counts from sites in the network, the in-depth work to select potential research subjects for recruitment is performed at each local site using informatics tools that we have developed.

Methods & Results
At Partners Healthcare, we have developed an i2b2 plugin that enables the viewing and exporting of aggregated patient data—one patient per row—from an i2b2 data warehouse. The plugin makes use of a new scalable processing engine that we developed called ai2b2 which assists with constructing the output file from the i2b2 Patient Data Object (PDO) in a multi-threaded and shared worker pool approach. This enables the ability for local investigators to review and retrieve aggregated patient data files from i2b2 in a reliable, efficient, and self-service way.

The investigator first picks a query in i2b2 that represents the desired patient population for the output file, chooses health-related criteria—such as demographics, diagnoses, medications, or labs—to define the columns of the file, and finally, selects the desired aggregations to generate. This workflow produces a “job file” which the ai2b2 engine utilizes to start spawning workers in the background to create the file (Figure 1). The result of this plugin is a comma separated value (CSV) file constructed from many “blocks” created by each worker communicating with i2b2—completely in the background and detached from the user’s browser session.

![Workflow Diagram](image)

**Figure 1.** Workflow after an investigator starts the file creation process from the plugin

Conclusion
The scalable process to be able to generate high-quality and relevant data sets from the i2b2 data warehouse has many applications. Existing i2b2 data warehouses already support the capability to serve patient data beyond the aggregate count. However, this plugin helps to unlock, organize, and present that data to an investigator in a managed way. For the ACT project, this plugin supports the patient identification workflow by producing aggregated data sets which investigators can use to help find potential clinical trial participants.
Great expectations: patient response to portal invitations to contribute to their electronic medical records

Emily C. Webber, MD1,2,3; Courtney Hines2; Leiza Schmoll2; Danyelle Binko, MHA2; Jason T. Schaffer, MD2,3; and Seung Park, MD2
1Riley Children’s Health, Indianapolis, IN; 2Indiana University Health, Indianapolis, IN; 3Regenstrief Institute, Indianapolis, Indiana

Introduction Patient interest and ability to use a portal to their EHR can be variable, despite widespread use1 and aligning patient and clinician needs with technical functionality is critical to success.2 In this project, we report a project to deploy patient contribution forms across several hundred clinics, allowing to patients to contribute data to their EHR through their portal, employing continuous improvement to sustain high levels of responses.

Methods Content for “Clipboards” was selected for widespread value to clinicians and patients, including analysis of clinical data and questions routinely gathered on paper forms. The primary components included: medical history (diagnoses or problems), procedural history, allergies, and a review of systems. The type of visit most likely to collect this data were included, and mapped to an appointment type. The “Clipboards” were distributed via portal to patients with the upcoming appointments, and the patients received an email alerting them to a new form in their portal and reminding them of their upcoming appointment. After completion and being submitted by the patient, the data was reviewed during the visit by the clinical team for inclusion or rejection from the EHR. There were two changes during our study: the first was an update to the “Clipboard” content which removed the review of system and reduced the overall number of Clipboards (based on patient and clinician input). In the second change, patients did not receive e-mail notifications of a waiting Clipboard.

Results Patient response rate in the first month increased to around 28% consistently. The review of systems was noted to be ‘too technical’ by patients leaving comments on the portal itself. Initial feedback from specialists was that the included components were not always addressed, particularly the review of systems section. After removing some appointment types and replacing the review of systems with a free text field for “concerns this visit”, patients continued to complete the Clipboards around 28% of the time, with a smaller group of appointments included. After the removal of the email notification reminders, patient response remained consistent.

Discussion Patient portal efficacy has been described as a means for niche populations or disease specific efforts to improve engagement; however, in this effort we describe a systematic effort to leverage the portal to for common elements of medical history. Response rates by patients have remained stable at 28%. We anticipate that with additional content that is widely appealing – including the medication list – will increase this response rate and establish new expectations for a ‘normal’ rate of patient engagement with this particular element.

References
Introduction: Prescription opioid use has more than tripled since the early 2000s and accounts for approximately 24 percent of all opioid misuse-related overdose deaths in the United States in 2017. Previous studies have relied on structured data (International Classification of Diseases - ICD codes) to estimate the prevalence of opioid-related utilization. Concerningly, without including unstructured clinical data these studies may have underestimated the true prevalence of opioid-related utilization. The purpose of this study is to compare prevalence estimates of opioid-related emergency department (ED) encounters using only structured data to a combination of structured and unstructured data using a natural language processing (NLP) approach.

Methods: Our sample included ED encounters from two hospital systems in Indiana from 2012-2017. Patient demographic and encounter data, 30-day prescription history, and unstructured clinical notes were drawn from a statewide health information exchange. We estimated the prevalence of opioid-related ED encounters to analyze trends in encounters using structured data, unstructured data, and a combination of both. To extract narrative clinical text, we developed a dictionary of opioid-related terms appropriate for the ED setting and comparable to past research. A rule-based NLP approach was used to identify encounters that included an opioid term (e.g. “heroin”, “oxycodeone”) in combination with a problem use term (e.g., “abuse”, “misuse”) from unstructured clinical text. We conducted a preliminary computer-assisted chart review of the NLP search strategy on a sample of each cohort (ICD, NLP, NLP/ICD). Kappa coefficients were produced to assess agreement between two abstractors among the three identification approaches. Lastly, we estimated a logistic regression model to identify patient, encounter, community, and prescription history associated with being ‘miscoded.’ We considered an encounter to be miscoded if NLP identified an encounter as opioid-related but there were no opioid misuse-related ICD codes identified in that encounter.

Results: A total of 13,352 patient encounters were identified as having been opioid related using only structured data, with significant increases from 2012 to 2017 (Figure 1). The rate of ICD code-identified encounters was 57.4 per 10,000 in 2017. When we included encounters identified using NLP, the rate of overall opioid encounters increased to 132.6 per 10,000 prior to validating the NLP approach. Inter-rater reliability between two abstractors was notably high for ICD-identified encounters (κ=0.87, and 91% agreement), NLP-identified encounters (κ=0.89, and 98% agreement), and a combination of NLP- and ICD-identified encounters (κ=0.90, and 100% agreement). Although agreement between abstractors was high, the NLP algorithm misidentified opioid misuse 48% of the time in the preliminary review. We expect the true prevalence rate of NLP-identified opioid misuse encounters to be approximately 95 per 10,000 given the false positive rate of the NLP algorithm. Regression results indicate that daytime encounters (OR: 1.213, p<0.003) and Medicaid beneficiaries (OR: 1.024, p<0.001) have greater odds of a miscoded opioid-related encounter.

Conclusions: This research reveals a gross underestimation of the prevalence of opioid misuse encounters in the ED. Natural language processing methods may provide better prevalence estimates of opioid misuse encounters when combined with structured data; however, reassessing predetermined rules may yield more accurate estimates. Encounter and patient characteristics also indicate whether an encounter is miscoded using NLP methods versus ICD codes, which may further inform the need to retest NLP rules. Additionally, future research will utilize comparisons of machine learning based lexical identification methods.

References:
Exploration of the Dental Informatics Working Group in the American Medical Informatics Association

Karmen S. Williams, DrPH, MBA¹, Neel Shimpi, BDS, MM, PhD²
¹Indiana University Fairbanks School of Public Health & Regenstrief Institute, Indianapolis, IN, USA; ²Center for Oral and Systemic Health, Marshfield Clinic Research Institute, Marshfield, WI, USA

Abstract

Dental informatics, as a potential core discipline, is yet in its infancy. American Medical Informatics Association’s (AMIA) Dental Informatics Working Group is comprised of various interdisciplinary practice, academic, and researcher leaders. To understand the insight of practitioners and researchers in the field, an 8-question survey was sent to the membership of the Dental Informatics Working Group. The perspectives of this survey were used to develop goals and objectives and to build action plans for implementing a positive change and active participation in the working group.

Introduction

New and still emerging field within informatics, dental informatics has the potential to bridge the gap between clinical care coordination in dental and medical settings. As an even newer working group within AMIA, dental informatics seeks to serve, collaborate, and share information within and outside AMIA communities. The mission of the Dental Informatics Working Group (DI-WG) is: to serve individuals and organizations with interest or involvement in health informatics with an emphasis on dental informatics; to collaborate with other health care specialties to improve healthcare applications; and to share expertise and experiences in the use of dental information systems, computer technology, and communications to improve patient care, research, and education. The objective of this project was to gain perspective of the DI-WG membership and the impact of DI-WG on the field of dentistry, dental informatics, and dental research.

Methods

A voluntary and anonymous, 8-question online survey of multiple choice and open text questions, was sent to the members of the DIWG for a duration of 4 weeks prior to the 2018 AMIA Annual Meeting. The questions included an assessment of years involved in dental informatics, research interest areas related to applying informatics to dentistry, challenges and opportunities in dental informatics, inciting the next generation, potential presentation topics, interest in presenting topic, and additional comments on increasing DI-WG presence in AMIA.

Results

A response rate of 19% (10/52) was achieved. Based on the summary of 52 members, 33 members represent academic organizations, 12 hospitals and industry, and 7 research institutions. Approximately 30% of respondents were been involved in dental informatics for 2-5 years, and 10 percent with over 20 years of experience. While one person was not involved in research, clinical decision support systems and tools, quality improvement, and dental and medical data integration was the most commonly mentioned research interest. Multiple challenges were indicated including linkage of dental records to rich physical health data, removing siloes and working together with meaningful collaboration among dental informatics professionals, and training and awareness of informatics in dentistry. Targeted educational curriculum, research opportunities being made available to students, and systematic outreach and recruitment strategy were among the responses for inciting the next generation.

Conclusions

Dental informatics is a building field within health informatics with an exciting outlook on growth for the future. Collaboration among dental informaticians and health care was a reoccurring theme. The DI-WG has a commitment to serve, collaborate, and share with all health specialties for the ultimate goal of true patient-centered care.
Automated Discovery of Common Data Elements in HIV Clinical Trials
Using an Interaction Network
Nick D. Williams, Ph.D, Craig S. Mayer, MS, Sigfried Gold, MFA, MA,
Vojtech Huser, MD, PhD, Kin Wah Fung, MD
1Lister Hill Center at the National Library of Medicine, Bethesda, MD, USA

Introduction
HIV clinical trials represent an untapped data re-use opportunity in a crucial research area. Reusing HIV clinical trial data variables in secondary analysis sourced from multiple studies requires the discovery of common data elements (CDE) across HIV studies. Ideally, clinical trials should re-use CDEs right from the beginning when the case report forms are designed. But in reality, most trials invent their own data elements and finding commonality across trials is only an afterthought. Very often, the burden of finding CDEs falls on the data analysts who re-use the data.

Methods
We collected clinical trial variable descriptions from 15 HIV clinical trial data dictionaries. Trial dictionaries were sourced from several harmonization schemes including dbGaP, Veterans Affairs, BioLINCC, IeDEA and NIDA DataShare. Extracted variable descriptions were mapped to UMLS concepts through MetaMap. Eleven semantic types (e.g., birds and plants) were excluded empirically; as they returned no relevant UMLS concepts. Variable description-UMLS concept pairs were processed in Cytoscape into a UMLS concept (node) to variable description (node) interaction network. The network was clustered using Mcode; which was tuned to produce the largest number of clusters; not the most UMLS concept specific. All identified clusters were reviewed manually.

Results
A total of 4,867 variables mapped to 4,282 unique UMLS concepts. The network contained 9,149 nodes and 29,450 edges. Altogether, 251 clusters encompassing 1,758 variables were identified with Mcode. Manual review confirmed 55% of the clustered variables were true positives (TP) i.e., the variables were the same or closely related to their UMLS concepts. Many of the false positive variables (FP) were cases in which variables which should form two or three clusters were mingled (Table 1). If an algorithm could be trained to correctly detect the cluster boundary in the bi-centered and tri-centered clusters, the overall true positive rate would become 88.8%. We found 566 TP variables (in 77 correct clusters) that originated from two or more distinct studies. These would be the likely variables which could become CDEs across studies.

Table 1. Variable Description True and False Positives in Result Clusters

<table>
<thead>
<tr>
<th></th>
<th>Variable TP</th>
<th>Variable FP</th>
<th>% of TP clustered variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct clusters</td>
<td>1,091</td>
<td>84</td>
<td>55.13%</td>
</tr>
<tr>
<td>Total</td>
<td>1,175</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusions
Network clustering method using MetaMap annotations shows promise for automating CDE discovery for clinical trial variables in HIV studies.

Acknowledgement
This work was supported in part by the Intramural Research Program of the National Library of Medicine and the Office of AIDS Research, NIH.

References
Discovery of Nurse-Patient Assignment in Medication Dispensing Data
Dana M. Womack, PhD, RN, Michelle R. Hribar, PhD, Paul N. Gorman, MD
Oregon Health & Science University, Portland, OR, USA

Introduction
Clinician burnout is an important workplace concern that is attributable to multiple causes. In particular, excessive workload is a predictor of nurses’ intention to leave their organization and profession, and it is a contributor to missed nursing care. In applied informatics research, the nurse-patient assignment is an important unit of analysis because it reflects nurse workload. Digital availability of nurse-patient assignment is often lacking, as assignment data is often recorded on paper in hospitals that utilize electronic health records. To support research in hospitals that lack electronic assignment data, there is a need for automated derivation of nurse-patient assignment data from available electronic data sources. Medication dispensing data includes information about the nurse, patient, and medication administered, making it a potential source for the nurse-patient assignment.

Purpose
The purpose of this study is to develop and validate a method for automated discovery of nurse-patient pairs using data produced by medication dispensing cabinets.

Methods
The Cross-Industry Standard Process for Data Mining was employed to develop an understanding of patient assignment practices and data sources, prepare a data set, define a method, and evaluate method accuracy. Data were extracted from an Omnicell® automated medication dispensing cabinet. The nurse who dispensed the greatest number of medications for a patient during a work shift was identified as the patient’s assigned nurse. Method accuracy was assessed through comparison of derived and paper-based pairs for a 1-month time period. Discrepancies were reviewed by a Technical Expert Panel which provided recommendations for interpretation of observed differences.

Results
A total of 1,400 nurse-patient pairs were derived through application of the method to data from a 1-year time period. In the 1-month test period, paper sheets contained 10 nurse-patient pairs not present in the derived set, and the derived set produced 18 pairs that were not present on paper. Derived nurse-patient pairs matched paper-based pairs in 393 of 450 test cases (87% accuracy) using exact match criteria.

Discussion & Limitations
Medication dispensing data contains information that can determine nurse-patient assignments with reasonable accuracy compared to paper records of nurse-patient assignments. Paper assignment sheets were not a complete record since some patients who received medications during a shift did not appear on the assignment sheet. The technical expert panel advised that missing patients likely represent new admissions, and suggested that derived data may more accurately represent historical data. With that assumption, accuracy of derived assignments increases to 91% as compared to paper assignment sheets that were 96% complete. A limitation of this method is that it works for historical data only. If used in real-time, accurate assignments would not be available until nurses have administered a majority of scheduled medications during a shift. In addition, patients who do not receive medications are not observable.

Conclusion
Retrospective derivation of nurse-patient assignment data is feasible in hospitals lacking electronic assignment data. The defined method produces high accuracy and serves as a facilitator for research at the nurse assignment level.

References
Recommendations for a Medical Device Knowledge Base

Melinda Wong, BS Pharm, RPh,1 Karen M. Bavuso, RN, MSN,1
Emily Man Wai Chan, PharmD, RPh,1 Martin Schotel,1 Saverio Maviglia MD, MS,1,2,3
Roberto A. Rocha, MD, PhD1,2,3
1Semedy, Inc., Waltham, MA; 2Brigham and Women’s Hospital, Boston, MA;
3Harvard Medical School, Boston, MA

Introduction
Medical devices are essential for the diagnosis, management, treatment, or prevention of disease or condition. Over the past few years, there has been significant global legislative and regulatory changes in the medical device industry in an effort to establish standardization and tracking of medical devices for safety and performance. However, there is a lack of knowledge base with clinical device information which can enhance quality of care and improve clinical outcomes (e.g. find medical devices manufactured with latex).1 We provide recommendations for types of information to include in a medical device knowledge base to support clinical decision support systems.

Methods
We reviewed publications and other online sources to identify potential information gaps pertaining to the representation of medical devices and creation of an integrated knowledge base.1,2 We assessed adverse events associated with medical devices from a published case report.3 In addition, we reviewed event descriptions of a subset of class II and III devices (i.e. categorized as moderate to high risk to patients and/or users) from the adverse event database managed by the “Food Drug Administration” (FDA), known as the “Manufacturer and User Facility Device Experience” (MAUDE).4 There were limitations in reviewing these sources since the narrative descriptions were based on observations reported and not systematically collected using a predefined protocol. However, for the purpose of knowledge representation, these sources provided general understanding and opportunities for developing a medical device knowledge base.5

Results
We successfully identified several properties that should be included in a medical device knowledge base: (a) product identifiers and descriptors (e.g. device identifier, model number, catalog number, serial number, expiration date); (b) use materials that can cause hypersensitivity reactions (e.g. natural rubber latex, cobalt alloy); (c) possible device interactions including device-device and device-human tissue (e.g. wearable defibrillator and ventricular pacemaker; surgical stapler and human tissue); (d) storage and handling requirements (e.g. specific temperature and humidity levels); (e) requirements prior to use (e.g. sterilization requirements); (f) warnings, precautions, and adverse events; (g) maintenance requirements (e.g. calibration and testing); and also (h) indications for use.

Conclusion
A medical device knowledge base can provide a valuable resource to health care providers, while also aiding in the decision-making processes. We confirmed that adverse events from medical devices and survey responses from health care specialists can provide insights on the type of information that must be included in a medical device knowledge base, contributing to patient safety and outcomes.

References

1817
Title: Identifying documented medical non-adherence from clinical notes using natural language processing

Joy Tzung-yu Wu, MD MPH\(^1\), David William Grant, MD\(^2\), Shrey Lakhotia, BDS\(^3\), Patrick D Tyler, MD\(^4\), Daniel F Gruhl, PhD\(^5\), Chaitanya Shivade, PhD\(^6\), Sebastian Gehrmann, PhD\(^7\), Leo A Celi MD, MS MPH\(^8\)

\(^1\)Pulmonary and Critical Care Medicine, Beth Israel Deaconess Medical Center, Boston, MA; \(^2\)Plastic Surgery, Barnes Jewish Hospital, St. Louis, MO; \(^3\)IBM Research - Almaden, San Jose, CA; \(^4\)T.H. Chan Harvard School of Public Health; \(^5\)Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA; \(^6\)John A. Paulson School Of Engineering And Applied Sciences, Harvard University, Cambridge, MA

Introduction

Patient non-adherence (NA), which is estimated to cost around $100 billion annually to the United States healthcare system is an example of clinical information that is only recorded in unstructured clinical notes\(^1\). NA is a behavior when patients do not follow important medical advice. It is thought that NA increases the risk of undiagnosed and poorly-treated illnesses, with consequences ranging from worse patient-centered outcomes to biasing results of clinical drug trials if the effect was not properly controlled.

Our contribution is twofold. Firstly, we present an experts-in-the-loop distantly-supervised NLP pipeline to extract NA. Note that this is documented non-adherence, which results from clinicians choosing to make a note. Such decisions may be shaded by clinician bias but this analysis of this is out of scope. The NLP pipeline was developed without spending significant clinical resources on annotation (<30 hours). We show a comparison of the estimated clinical resource if we had tried to develop a fully supervised NLP algorithm to extract the NA feature. Our second contribution is a quantitative characterisation of patients exhibiting “documented non-adherence”. To the best of our knowledge, this has not been done before in the clinical literature. The patient population in our case came from a tertiary academic hospital in the United States. However, the approach presented should be replicable to extract the NA feature and understand characteristics of other patient populations.

Methods

NLP pipeline - Clinicians used an unsupervised “salient terms” proposal and curation NLP tool\(^2\), which has a user-friendly interface, to efficiently expand the vocabularies indicative of NA if mentioned in clinical notes. We also clinically grouped these documented NA terms into 5 subtypes, which include: 1) medication NA, 2) appointment NA, 3) refusal against medical advice, 4) dietary NA, and 5) NA, not otherwise specified. Subsequently, a rule-based NLP pipeline was developed to extract the NA features from notes. First we detect whether a NA term was mentioned by a note. If no mention, the note is negative for NA. If mentioned, we used NLTK to tokenize sentences in the note, and Negex to detect negation at sentence levels. If NA was not negated in most sentences in a note, then the note is deemed to contain evidence of documented NA. To improve recall, medication and dietary NA were broken down to their conceptual components (medication, diet, and non-adherence) to be captured separately if they occur in the same sentence (e.g. patient [missed] several doses of her [warfarin]). Subsequently, any hospital admission that includes any note with evidence of documented NA as per our NLP pipeline was predicted to have NA.

Test dataset for agreement - Two clinicians (DWG, SL) independently annotated 300 notes (50 notes randomly sampled from notes that just mention terms for each predicted NA subtype plus 50 notes that do not mention any NA terms) to evaluate the NLP pipeline.

Clinical analysis - An open source intensive care unit (ICU) database (MIMIC-III\(^9\)) is used for this study. We included all patients aged between 10 and 90 at their first hospital and ICU admission, and performed a quantitative analysis of their demographic, comorbidities, ICU severity, readmissions and mortality outcomes as grouped by the presence or absence of NA and its subtypes.

Results: Table 1 shows the inter-annotator and agreement with the NLP results on NA overall and for each specific NA subtype.

<table>
<thead>
<tr>
<th></th>
<th>NA overall</th>
<th>Medication</th>
<th>Appointment</th>
<th>Dietary</th>
<th>Refusal AMA</th>
<th>NA, NOS</th>
<th>No evidence of NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-annotator % agreement</td>
<td>0.88</td>
<td>0.84</td>
<td>0.76</td>
<td>0.94</td>
<td>0.88</td>
<td>0.74</td>
<td>0.92</td>
</tr>
<tr>
<td>Average NLP % agreement</td>
<td>0.87</td>
<td>0.84</td>
<td>0.74</td>
<td>0.89</td>
<td>0.80</td>
<td>0.79</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Conclusion

We curated lexicons of terms indicating five different different types of documented NA behaviors in notes. Using these lexicons we built a rule-based NLP pipeline to identify NA patients. An evaluation of the pipeline using gold standard curated by clinicians resulted in a performance of 87% accuracy. A comparison of NA across different dimensions in the EHR shows interesting trends.

References


1818
Development of a Consult Management Tool to Streamline Consult Scheduling Process and Improve Staff Satisfaction

Jianji J. Yang, PHD, Jonathan Sun, MS, Mika Asher, BS, Jerry Kohler
VA Portland Healthcare System, Portland, Oregon

Introduction

Electronic Consult is a common feature in EHR systems, such as Computerized Patient Record System (CPRS). In CPRS, a consult is used as a two-way communication on behalf of a patient consisting of a provider (sender) request seeking expertise from another provider (receiver) regarding evaluation or management of a specific problem. Upon receiving the consult, the receiving service will schedule a visit with the patient to evaluate the issue before providing expert advice to the sender.

In 2016, Veterans Health Administration (VHA) issued a Directive to update and standardize consult business rules, processes and procedures, clearly defining oversight responsibilities. It outlines the requirements and timelines for each steps of the consult processes in details, e.g. submission, unit level management, tracking and closure.

Problem Statement

The Directive requires action taken within 3 days after a consult is submitted and multiple scheduling attempts (phone calls and letters) over a period of 2 weeks prior to discontinuing a consult if no response from the veteran. There was no system to track the consult lifecycle, and no good ways for staff to check where they were and where they had left off. Because of the asynchronous nature of scheduling, schedulers relied on a labor-intensive manual approach involving daily printing of lists for review in CPRS to determine action needed for each consult. The process resulted in duplication of work, delays in timely contacts and consult discontinuation, as well as low staff satisfaction. The aim of this project was to explore an informatics approach to improve efficiency of the proces, and as a result, improve staff satisfaction and veterans timely access to care.

Informatics Tool Development

The VAPORHCS informatics team developed a web-based Consult Tracker application to support timely scheduling activities. The development utilized human-centered user-interface design, agile development process and usability evaluation to inform enhancements. The application leverages data from VA’s Corporate Data Warehouse and employs a service-oriented architecture with multi-layer reusable modules and data services. Its user-interface provides a clear situation awareness for all consults in scheduling phase, as well as day-to-day action filters e.g. 1. Displays all actions due today, 2. Allows filtering by consult priority, 3. Provides high visibility of STAT consults.

Current Status

After piloting in three specialty services and additional usability enhancements from pilot feedback, the Consult Tracker is currently used in all medical and surgical services. Services reported improved staff satisfaction, more timely contacts and discontinuation of consults, and reduction of STAT consult pending time. It helps to streamline consult scheduling process, improves workflow consistency across services, and eventually might have a positive impact on veteran access to care. Scheduling is the first step of the consult process; additional informatics support is needed for other stages of this complex asynchronous communication feature to work efficiently and reduce delay.

Discussions

VAPORHCS developed a web-based application to support the scheduling effort for consult management. It helps to improve efficiency and meet the VHA Directive. There are much informatic efforts to assist clinicians in their care of patients, schedulers and other support staff are integral and essential components of healthcare delivery. Supporting their workflow with informatics tool in turn improves efficiency of healthcare system, patient access and outcome.
Do People with Diabetes Receive Balanced Diabetes Self-Management Education? An Analysis of Clinic Notes against AADE7™ Principles
Qing Ye, MS¹,², Richa Patel, MD³, Uzma Khan, MD³, Suzanne A. Boren, PhD, MHA¹,², Min Soon Kim, PhD¹,²
¹Informatics Institute; ²Department of Health Management and Informatics; ³Department of Medicine, University of Missouri, Columbia, MO

Introduction
Diabetes self-management education and support (DSME/S) delivered by medical providers has been shown to improve patient engagement in self-care. However, the time providers spend with their patients is limited, and may limit the delivery of DSME/S efficiently. The American Association of Diabetes Educators Self-Care Behaviors™ (AADE7™) provides structured and validated diabetes self-management guidelines to help providers deliver the seven key principles of DSME/S. The objective of this study was to use AADE7™ guidelines to identify gaps in the knowledge presented by physicians during clinic visits. The research question was “Do the clinic notes address the seven principles of AADE7™?”

Methods
We used PowerInsight®, a clinical reporting platform for electronic health records, to identify people with type 1 or type 2 diabetes, only listing those who were 18 years or older and who visited the Diabetes and Endocrinology Center at the University of Missouri Health Care between January 1, 2017 and January 1, 2018. We applied exclusion criteria, such as pregnancy, incarceration, chronic liver or kidney disease, or HIV to create a patient list of 100 patients. For each patient, we extracted History of Present Illness (HPI) and Impression and Plan (I&P) sections from two clinic notes per patient. In these two sections, providers record patients’ daily self-management behaviors and provide recommendations for their patients. We developed a codebook by consulting the AADE7™ guidelines for the education items on Healthy Eating (HE), Being Active (BA), Monitoring (M), Taking Medication (TM), Problem Solving (PS), Reducing Risks (RR), and Healthy Coping (HC). We employed a multi-step coding process over a 10-week period along with bi-weekly group reviews. Ye and Patel coded the clinic notes based on the codebook. An interdisciplinary team (Khan, Boren, Kim) randomly selected and reviewed 10% of the clinic notes to ensure accuracy of coding. Using deductive thematic analysis, we first evaluated if the particular AADE7™ principle was addressed and then counted how many times it was addressed in each patient’s two clinic notes. We used the Generalized Linear Mixed Model (GLMM) to predict the most likely outcome for each of the seven principles of the AADE7™ for each patient. For each of the seven principles, we used the Mann-Whitney tests to understand whether the count of codes differed based on sex and age.

Results
The codebook included HE (12 codes), BA (13 codes), M (17 codes), TM (13 codes), PS (6 codes), RR (9 codes), and HC (8 codes). Out of 1,317 patients, 100 patients were selected, resulting in 200 notes with 400 note sections for coding. There were 68 patients aged 18-64 years and 32 patients aged 65 years and older. There were 52 males and 48 females. The count of codes for each principle were divided by the number of clinic notes. In HPI section, the most common principles were M (mean=2.610, median=2.500) and TM (mean=1.570, median=1.500). The least common principles were BA (mean=0.245, median=0) and HC (mean=0.270, median=0). In I&P section, the most common principles were M (mean=6.430, median=6.750) and TM (mean=1.830, median=2.000). The least common principles were HC (mean=0.105, median=0) and BA (mean=0.320, median=0). The odds of appearance were defined as the ratio probability of appearance over the probability of no appearance. In HPI section, the GLMM showed that the odds of the M appearance divided by the odds of the BA appearance was 27.315. In I&P section, the GLMM showed that the odds of the M appearance divided by the odds of the HC appearance was 72.257. That is to say M had a higher probability of appearance in both HPI and I&P sections. Mann-Whitney tests indicated that there were no statistical differences in the count of codes in the different sex groups and age groups in each of the seven principles.

Conclusion
This study showed clinic notes did not address each of the AADE7™ principles equally, which suggests that patients may not receive balanced DSME during their visits. This unbalanced approach by providers has implications for long-term diabetes management. One of the reasons why information about Monitoring occurred most in the clinic notes may be because of the health insurance requirement. For example, current Medicare requires that providers should document information about blood glucose data and HbA1c to have glucose testing equipment and supplies covered for their patients. Lack of discussion about Being Active and Healthy Coping therefore may have long-term implications because obesity and depression are common comorbid conditions in people with diabetes. Further studies are needed to identify the other underlying reasons for the imbalance.
Visualizer for Evaluating Adherence to Evidence-Based Guidelines

Jingzhi Yu, BA¹, Farhad Ghamsari, BS, MS¹, Samuel Ross, BA¹, Vesna Mitrovic, MS¹, Abel Kho, MD, MS¹²

¹Center for Health Information Partnerships, Feinberg School of Medicine, Chicago, IL; ²Department of Medicine, Northwestern Memorial Hospital, Chicago, IL

Abstract

For many small providers, integrating and visualizing information from multiple sources within their EHR could be difficult given limited training and financial resources. We have created a desktop-based patient history visualizer that takes in patient C-CDA files and easily produces a visualizer for measuring adherence to evidence-based guidelines on co-morbid diabetes and hypertension. This visualizer will be piloted within a group of eligible small providers and the usage and effectiveness of this tool will be evaluated.

Introduction

With the HITECH Act of 2009, Medicare and Medicaid providers were incentivized to adopt and achieve “meaningful use” of Certified Electronic Health Record Technology (CEHRT) through implementation of key functionality and the capture and reporting of core clinical information. This effort was carried forward under MACRA in 2015, with an increased focus on improved outcomes at the lowest cost¹². In order to succeed in rapidly advancing value-based payment models, providers need tools that organize troves of data relevant to quality of care, which can be a complex and time consuming process. Currently available tools providing such functions are more text-based than intuitive visual-based, for physicians and patients alike³. Evaluating adherence to evidence-based guidelines is one type of complex process that requires integrating different sources within CEHRT. Electronic clinical quality measure (eCQM) reports offer performance scores and lists of at-risk patients but do not link to actionable decision support for closing gaps in treatment protocols. During the process of assisting providers with implementing evidence-based care plans, the need for a more convenient and customizable tool became evident.

Methods & Results

We are designing a patient history visualizer that is desktop based, EHR-independent, and easy to use. End users only have to “drag and drop” their desired patient C-CDA files (extracted from CEHRT) to a designated folder in order to produce visualizations of adherence to evidence-based guidelines with the click of a button. C-CDA files are parsed using the glob and lxml package in Python. Patient encounters, procedures, medication and other types of clinically relevant information are extracted from the human readable table section of the C-CDA file.

Given the overwhelming volume of data and user interfaces experienced in navigating CEHRT, organizing a narrow set of information relevant to condition-specific goals is resource intensive. In order to develop a tool that addresses provider needs without manual chart abstraction, we are interviewing care teams focused on patients with co-morbid diabetes and hypertension. These chronic conditions were chosen due to their prevalence in the general population and priority as targets for quality measurement in value-based payment programs. Efforts to produce a visual evaluation of adherence to evidence-based recommendations in this population will serve as the blueprint for supporting the implementation of care plans in other use cases. Finally, we will document physician’s use of the tool and its impact on their performance on quality measures supported by value-based payment programs.

Conclusion

The results of this study could demonstrate proof of concept and benefits of our tool. This study draws attention to designing tools for providers that focus on ease-of-use and decision support with adherence to evidence-based protocols. In addition, more analytical tools should originate from the provider’s need, featuring customizable visualizations representing quality of care. This will not only allow the physicians to maximize the utility of CEHRT but also empower patients to realize the value of the data generated in the course of their care.

References

Participatory Design of a Web-Based Intervention for Parents to Improve Sleep in Young Children with Arthritis

Weichao Yuwen, PhD RN1, Uba Backonja, PhD MS RN1,2, Maggie H. Bromberg, PhD1, Natalie Donahue, BS RN2,3, Michelle M. Garrison, PhD MPH2,3, Sarah Ringold, MD MS2,3, Teresa M. Ward, PhD RN FAAN2

1University of Washington Tacoma, Tacoma, WA, USA; 2Universtiy of Washington, Seattle, WA; 3Seattle Children’s Hospital, Seattle, WA; 4Seattle Pediatric Psychology, Seattle, WA

Introduction

An estimated 25-40% of young children (2-5 years) in the United States have behavioral sleep problems (e.g., frequent night awakenings, requiring parental presence at bedtime). Sleep problems are even more prevalent among children with chronic health conditions such as juvenile idiopathic arthritis (JIA). Sleep problems in early childhood persist into middle childhood, and predict later development of emotional and behavioral problems, poorer quality of life, and increased health care needs. Face-to-face interventions have been developed to promote sleep health in healthy infants and young children. However, there are numerous barriers to access, including a shortage of trained pediatric providers in managing sleep problems, long waiting lists, transportation constraints, and costs. Additionally, children with chronic conditions are already frequently engaged with the healthcare system so may particularly benefit from distance-based interventions that do not require additional clinic visits. The use of technology to reduce these barriers and support families of children with chronic conditions who struggle with sleep is in its infancy. The purpose of this study was to use participatory design (PD) to engage parents and healthcare providers to elicit JIA-specific issues and needs that impact children’s sleep and in the adaption of an existing healthy sleep intervention to a technology-based format to meet the needs of this clinical population.

Methods

We recruited nine parents of young children with JIA and five healthcare providers from a pediatric rheumatology clinic at a children’s hospital in the northwest. We conducted two in-person PD sessions, one 2-hour session with five parents and one 1-hour session with five providers, respectively. Four parents who were unable to attend an in-person session (e.g., geographic constraints) were mailed a design packet and craft materials. The packet included prompts and the series of activities done in the in-person session. Parents then mailed back the completed packet. Example questions include: “What are things that you and your child do in order to get your child to bed on time? What are the challenges you and your child run into during the hour before bedtime?” Materials and audio recordings from the PD sessions were transcribed and analyzed using directed content analysis.

Results

Six parents and three providers were female, and 57% of the participants were non-Hispanic White. PD data was categorized into three domains: intervention content, features and functions, and other aspects. For content, participants shared common barriers to good sleep in children, sleep hygiene practices, and JIA-specific concerns and needs. Participants suggested many features and functions of a web-based intervention, such as aesthetic and minimalist design, symptoms tracking and data visualization, reminders, and peer connection. Healthcare providers also shared thoughts on how to design the intervention to be easily incorporated into the child’s clinical visits.

Conclusion

Engaging with the stakeholders and the end-users early in the development process through PD sessions and packets yielded valuable information that will be incorporated into our web-based intervention and empirically tested.

References

Plugin for importing Spreadsheets into Informatics for Integrating Biology and the Bedside platform

Akshay Zagade BE\(^1\), Vishal P. Vernekar BE\(^1\), Shek-Wayne Chan\(^4\), Kavishwar B. Waghoolikar MBBS PhD\(^2,3,4\), Rupendra Chulyadyo \(^3\), Yuri Ostrovsky PhD\(^1\), Alyssa P. Goodson \(^4\), Ameet Pathak MS\(^1\), Corey Glynn MS\(^3\), Christopher Herrick\(^4\), Shawn N. Murphy MD PhD\(^2,3,4\)

\(^1\) Persistent Systems, Pune, India and Santa Clara, CA; \(^2\) Harvard Medical School, Boston, MA; \(^3\) Massachusetts General Hospital, Boston, MA; \(^4\) Partners Healthcare Boston, MA; \(^5\) Brigham and Women’s Hospital, Boston, MA;

Introduction

The Informatics for Integrating Biology and the Bedside (i2b2) platform has been extensively deployed as a data warehouse to facilitate researchers in identifying patient cohorts for conducting research. However, the platform can be useful for individual researchers to query study/project-specific datasets. This is because study specific datasets can often be very large running into millions of rows, and hence cumbersome to query using spreadsheet software. The i2b2 web-client can readily provide the querying functionality for such datasets and assist non-expert users without knowledge of SQL or spreadsheet software to perform queries. To enable such a use case, we have developed a server-side plugin for i2b2, that will allow non-experts to easily import data from spreadsheets into i2b2.

Methods

The plugin is a daemon process that scans a file system folder for a ZIP file. This ZIP file contains a CSV (comma-separated values) file wherein each patient and that patient’s data are represented on a single row. On finding the CSV file, the plugin parses the first three rows for metadata, generates the i2b2 ontology, and then pivots the sheet to generate facts for insertion into the i2b2-fact observation table. The i2b2 ontology is created by inserting rows into concept dimension and metadata tables, such that the ontology is available for querying in the i2b2 web-client.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Date</th>
<th>Age</th>
<th>Zip Code</th>
<th>LDL</th>
<th>LDL Unit</th>
<th>LDL Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>src_patient_id</td>
<td>row_datetime</td>
<td>/Demo/Age</td>
<td>/Demo/ZipCode</td>
<td>/Labs/LDL</td>
<td>/Labs/LDL@Unit</td>
<td>/Labs/LDL@Date</td>
</tr>
<tr>
<td>src_patient_id</td>
<td>row_datetime</td>
<td>demo/a</td>
<td>demo/zip</td>
<td>labs:ldl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12345678</td>
<td>01/01/2019</td>
<td>56</td>
<td>01863</td>
<td>194</td>
<td>mg/dl</td>
<td>01/01/2019</td>
</tr>
</tbody>
</table>

The logic implemented for generation of the ontology from the first three rows is as follows (see Table 1): The first row contains a list of the column names. The second row contains the canonical concept paths for each item in the first row. The third row optionally contains the corresponding concept codes. Second row must contain a column with value ‘src_patient_id’ indicating that the column contains the patient id in the source system. Second row could optionally contain a column ‘row_datetime’, which is the time used for each fact generated for that row. If this column is absent the file modification date is taken as the start_date for each fact. Second row could optionally contain @date in the concept path, to indicate date modifier for the fact. Second row could also optionally have suffix of @unit will be stored as the unit for the fact. We benchmarked the plugin performance (Table 2), on 8 CPU machine (2.13GHz) with 32 GB RAM for the plugin, and 12 CPU machine (2.6 GHz) with 26 GB RAM for i2b2 containers.

<table>
<thead>
<tr>
<th>Index</th>
<th>Facts</th>
<th>Batch size</th>
<th>secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Indexes</td>
<td>30,000</td>
<td>1,000</td>
<td>32</td>
</tr>
<tr>
<td>Without Indexes</td>
<td>30,000</td>
<td>1,000</td>
<td>23</td>
</tr>
<tr>
<td>With Indexes</td>
<td>30,000</td>
<td>1,000</td>
<td>34</td>
</tr>
<tr>
<td>Without Indexes</td>
<td>30,000</td>
<td>1,000</td>
<td>37</td>
</tr>
<tr>
<td>With Indexes</td>
<td>300,000</td>
<td>1,000</td>
<td>310</td>
</tr>
<tr>
<td>Without Indexes</td>
<td>300,000</td>
<td>1,000</td>
<td>232</td>
</tr>
<tr>
<td>With Indexes</td>
<td>300,000</td>
<td>1,000</td>
<td>300</td>
</tr>
<tr>
<td>Without Indexes</td>
<td>300,000</td>
<td>5,000</td>
<td>270</td>
</tr>
<tr>
<td>With Indexes</td>
<td>3,000,000</td>
<td>1,000</td>
<td>3,000</td>
</tr>
<tr>
<td>Without Indexes</td>
<td>3,000,000</td>
<td>1,000</td>
<td>2,440</td>
</tr>
<tr>
<td>With Indexes</td>
<td>3,000,000</td>
<td>5,000</td>
<td>3,015</td>
</tr>
<tr>
<td>Without Indexes</td>
<td>3,000,000</td>
<td>5,000</td>
<td>2,520</td>
</tr>
</tbody>
</table>

Conclusion

Our benchmarks show that the speed of import is comparable to an existing excel import plugin developed by the BRISKSKIT project. We are currently investigating approaches to improve speed of the plugin.

References

Biomedical Research Infrastructure Software Service Kit, [https://www.brisskit.le.ac.uk/](https://www.brisskit.le.ac.uk/).
A natural language processing system to extract rich genetic and phenotypic information from Chinese clinical narratives

Xiaohui Zhang, MD, Yaoyun Zhang, PhD, Qin Zhang, PhD, Yuankai Ren, MS, Jianhui Ma, MB, Qi Sun, MD

1Peking Union Medical College Hospital, Beijing, China
2Digital China Health Technologies Co. Ltd, Beijing, China
3National Cancer Center, Chinese Academy of Medical Sciences, Beijing, China

Introduction

The prevalence and deaths of cancer cases in China stand at 21.8% and 27.1%, respectively, in consistency with those worldwide.[1] Notably, breast cancer ranks first among female, urging a rapid promotion of precision medicine. Fortunately, unstructured text in clinical notes of the electronic health records (EHRs) systems contain the richest phenotypic and genotypic test information, which can be extracted automatically by natural language processing (NLP) methods. Many efforts have been taken to extract such information in the US.[2] Despite that China has a large scale of EHR data from a representative population, the development of NLP-based tools and its applications for precision medicine is at its infant stage. This study takes the initiative to build NLP systems to facilitate the precision medicine of breast cancer.

Method

First, an information model was designed by domain experts, which contains a rich set of fine-granular genotypic and phenotypic information. In total, 58 entities (i.e., clinical concepts and their attributes) and 59 relations between them were defined (Ref. Table 1). Next, 100 encounter notes, pathology reports and procedural notes were manually annotated. The overall inter-annotator agreement was 0.81 for entities and 0.89 for relations. Then the current state-of-the-art BERT language model[3] pre-trained on a 3 Giga byte dataset of clinical text was fine-tuned, to develop the IE models for named entity recognition and relation extraction. Specifically, the named entity recognition was represented as a sequence labeling task and the relation extraction was represented as a binary-class classification task to determine whether two entities have a specific type of relation or not. The default settings of parameters as recommended in BERT were used for model training. [2]

Results & Discussion

Ten-fold cross validation was conducted for experimental evaluation. Overall, the IE system achieved a F-measure of 93.53% (68.97%~100%) for entity recognition and a F-measure of 91.4% for relation recognition (52.60%~100%). Given that a rich set of multiple types of entities (58) and relations (59) were recognized, the experimental performance was really encouraging and promising. Some entities and relations have low frequencies in the dataset (<10) and got the lowest performance. More data will be annotated in the future to increase the training set. Heuristic patterns and domain knowledge will also be used in the future to further improve the system performance.

Conclusions

Experimental results demonstrated the feasibility of using clinical NLP tools to extract rich genetic and phenotypic information from Chinese clinical narratives automatically. Typical use cases of precision medicine for breast cancer will be examined in the future leveraging information extracted from a larger scale of clinical data.

Reference

Challenges and Opportunities of Data Quality Improvement in Healthcare Organizations

Yili Zhang PhD, Urmita Banerjee, B'Tech, and Güneş Koru, PhD
University of Maryland, Baltimore County, Baltimore, MD

Introduction
With an increasingly volume of health data maintained in healthcare information systems, there are tremendous potentials to improve decision making and healthcare operations toward better care and health. Unfortunately, problems of data quality, which indicates whether data conforms the quality to serve the needs for specific goals, are wide spread in healthcare organizations. Therefore, systematically improving the quality of healthcare data in a continuous fashion is vital in healthcare organizations. Although there are a few research studies about data quality, how to improve data quality in a systematic and collaborative fashion in large healthcare organizations has not been explored yet. As the first step, this study contributes to the body knowledge by developing a comprehensive understanding of the challenges and opportunities toward collaborative and continuous data quality improvement in healthcare organizations.

Methods
Qualitative methods were adopted due to a need to conduct a deep investigation into this research in real-life settings of healthcare organizations. Upon being contacted by a State Health Department which expressed a need to understand the issues around data quality improvement, our study identified this department as the research context based on the resources made available for the project by the organization. Twelve semi-structured interviews were conducted to collect rich contextual data from the healthcare administration officials, who maintain and utilize of Medicaid data as their daily work. Established systems analysis strategies guided the formation of open-ended research questions. Expert sampling and convenience sampling were adopted to select participants who are in the position of data management in healthcare organizations and interact with Medicaid data in their daily work. The interviews were conducted over the phone, recorded, and transcribed verbatim to ensure accuracy. The data analysis of interview adopted Framework method, which is appropriate for interview transcripts in a systematic and flexible way. Respondent validation and member-checks are also taken to obtain accurate and reliable results. Data collection from interviews stopped when there is no new code found in data. Finally, a thematic framework to classify and organize data was generated.

Results
Three main themes identified as the challenges of data quality improvement are: (i) a lack of communication among healthcare professionals to develop a common understanding about data standards used in various systems, a participant mentioned: “You know I commonly ask for reports that I believe are clear but when I get the data it doesn’t answer my question”; (ii) the legacy system that manages Medicaid data without data validation features: “So if we can not completely upgrade the system, and the current system cannot take the improvement”; (iii) and a lack of documentation about data, which helps to understand the meaning, usage, correlation, and validation of each data item, as one said “If we don’t have good documentation, we may end up denying someone falsely”. The opportunities of data quality improvement are: (i) developing a data dictionary to describe features, standards, and uses of data items: “And I wish there would be a more prompt thing like how would this file be used and what do you want the end product to look like...”; and (ii) availability of automated tools to check for validity of data from different resources and groups: “... they receive updates from us once a week. Sometimes it will more beneficial if they can receive it everyday, but you know there are a lot of patients, sort of like that”. Participants realized problems in data quality when they claim a reimbursement, look at a specific record, or interact with external people such as providers. Their current solutions to improve data quality includes manual checks, various ad-hoc approaches, and using paper-based request forms.

Conclusion
The lack of data quality is considerably affecting the daily operations in the healthcare organization. However, currently employed practices fall short in effectively and efficiently improving the quality of data, and there is no systematic process yet in the organization to improve data quality according to participants. In the future, health organizations are likely to benefit from IT supported collaborative workflows that continuously improve data quality.

References
A Survey of Clinicians’ Perception of Pharmacogenomics Testing for Antidepressants

Yiye Zhang, PhD, MS1, Yonaka Harris2, George S. Alexopoulos, MD1; Adam Stracher, MD1; M. Elizabeth Ross, MD, PhD1, Rainu Kaulash, MD, MPH1, Jyotishman Pathak, PhD1
1Weill Cornell Medicine, New York, NY; 2New York City College of Technology, New York, NY

Introduction
Depression was identified by the WHO in 1990 as the fourth most disabling medical condition, and by the year 2020 it is predicted to be the second most disabling condition worldwide. (1) The current standard of care for depression is based on “trial-and-error” of a multitude of medications, including antidepressants. One way to reduce trial-and-error strategy is to use pharmacogenomics (PGx) testing to guide medication treatment and increase the likelihood of response by reducing the number of ineffective medication trials. While the scientific understanding of PGx is accelerating, its translation to routine clinical practice has progressed slowly. In an effort to investigate the barriers and opportunities in the adoption of PGx testing, we conducted a survey of clinicians’ current knowledge of and experience with PGx testing and their perception of its utility in routine clinical care.

Methods
A 41-question survey was created using Qualtrics (Qualtrics, Provo, UT). We distributed the survey electronically to 394 clinicians in Psychiatry, Internal Medicine, and Neurology at Weill Cornell Medicine (WCM) between October to November 2018. Participants were offered a $25 gift card as incentives to participate. All responses were anonymized and kept confidential. This study was approved by the WCM internal review board (IRB).

Results
As of March 2019, 69 surveys were started and 67 were completed (18% response rate). Results show that more young clinicians (under the age of 30, 31-40 and 41-50) had placed at least one order for PGx testing than physicians above age 50 (25% vs. 9%, Figure 1 left). Conversely, older physicians felt more well-informed about PGx testing for antidepressant than younger physicians (67% vs. 29% agree). Across specialty, 19 out of the 27 (70%) psychiatry providers felt well-informed about PGx testing for choosing antidepressant, whereas 3 out of the 21 (14%) internal medicine providers felt the same way (Figure 1 right). Furthermore, while a majority of the clinicians responded that they would prefer non-genetic testing over PGx testing for their patients in general (52%), internal medicine clinicians were the most likely to not order a PGx test for antidepressants (73% never ordered in the last year).

![Figure 1. Age vs. number of prescriptions, and specialty vs. feeling well-informed about PGx for antidepressant](image)

Conclusion
Findings from the survey suggest that while clinicians have knowledge of the PGx testing and understand the benefits, there still remain significant concerns and barriers in the adoption of PGx testing in routine clinical care.

Reference
Contextualizing Consumer Health Information Seeking

Zhan Zhang, PhD1, Caleb Wilson2, Zhe He, PhD2
1Pace University, New York, NY, USA; 2Florida State University, Tallahassee, FL, USA

Introduction

There is an ongoing trend for people turning to online resources, such as social Q&A sites and health forums, to seek personalized healthcare information. People are increasingly doing so in part due to many online health communities allowing them to ask full questions in natural language (as opposed to search terms) and to include key contextual health information with these questions. Prior work has highlighted the significant impact of contextual factors (e.g., people’s goals and motivations) on the result of information searching [1]. However, little attention has been paid to examining how patients formulated their questions to fulfill their information needs. Built upon our prior work [2], this study set out to examine the correlation between the types of information people provided and the types of support people were seeking by analyzing questions that people posted in a major social Q&A site.

Methods

Details on data collection and analysis has been published previously [2]. Briefly, two researchers independently performed content analysis on 967 randomly sampled posts containing laboratory test results pertaining to diabetes (e.g., HbA1c, glucose), collected from Yahoo! Answers between 2009 and 2014. The data analysis process consists of codebook development, group discussions, inter-rater reliability testing, and qualitative coding. This analysis helped us identify the types of information that people needed and provided. In this study, we further analyzed the posts identified through our previous work, and focused on the correlation between the types of information people provided and the types of support people sought. Figure 1 shows frequently observed correlations. The number represents the frequency where a particular type of information need and information provided co-occur in the posts.

Results

We found that people tended to provide different types of personal health information in a post, including demographic and medical information, to explain their current conditions to the community members to elicit relevant answers. We also found that specific types of information provided coincide with certain information needs. For example, when people sought help understanding the results of lab test, they often provided more details about the context of lab test they took (e.g., lab procedure, what they eat or drink) in addition to the results (e.g., “I drank sweet tea at 8:30 am and had my blood drawn at 12:30 pm”). Therefore, consumer-facing informatics tools should be designed, at both algorithm and interface levels, to present lab test results within patients’ contexts (e.g., age, medical history). Even though some contextual elements may not necessarily affect the accuracy of test results, the systems should provide explicit, consumer-friendly explanations to address any concerns or assumptions that patients may have. It is also worth noting that people would express their emotions when they asked for help. This observation suggests that informatics tools should be designed more emotion-responsive by providing means for patients to seek emotional support, i.e., through a social network in online communities to connect with peers who have similar conditions.

Finally, we want to note that many of the findings are generalizable to other chronic conditions even though the work was done in the context of diabetes. The results could have implications for the design of context-based consumer informatics tools, including EMR, to provide more personalized information that better meets patients’ needs.

References

Deep Learning Using Electronic Health Records and Genetic Data to Predict Cardiovascular Diseases

Juan Zhao, PhD1; QiPing Feng, PhD2; Patrick Wu, BS1; Joshua C. Denny, MD, MS1,3, Wei-Qi Wei, MD, PhD1

1Department of Biomedical Informatics, 2Division of Clinical Pharmacology, 3Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Introduction
Cardiovascular diseases (CVD) is a complex disease, which involves genetic, environmental, and lifestyle causes. Despite well-known prediction tools, such as the Framingham studies that identified several conventional risk factors (e.g. age, smoking and hypertension), CVD remains the leading causes of deaths and major global burden. Thus, early prediction for CVD is essential. Previous studies have identified genetic variants associated with CVD.1 Polygenic scores have been used to summarize genetic contributions for individuals’ traits, however; how to combine genetic variants with other biological and lifestyle factors remains a challenge. Deep learning models such as convolutional neural networks (CNN) and LSTM can explore temporal patterns on electronic health records (EHR) data to enhance disease prediction.2,3 In this study, we utilized a concatenating framework of CNN to combine temporal EHR data and genetic features for prediction. We compared the predictability of this approach with models using only EHR data.

Method
Data Description: The study used Vanderbilt University Medical Center’s (VUMC) de-identified EHRs and BioVU, the largest single-site biobank in the U.S. All individuals had 10-year follow up from 01/01/2007 to 12/31/2016. We extracted 7-year EHR records prior to 01/01/2007. We defined cases as individuals with ≥1 CVD diagnosis codes (ICD-9-CM: 411; * or 433; *) from 01/01/2007 to 12/31/2016; controls are individuals without any CVD diagnosis codes. We divided the observation period into one-year slice window to extract the temporal features. The study was limited to European and African ancestry adults.

Features: We extracted demographic data (i.e., age, gender, and race), lifestyle factors (i.e., smoking, BMI), lab and measurements such as systolic/diastolic blood pressures, total cholesterol, high-density lipoprotein, and preselected diagnosis and medication codes. We chose 204 single nucleotide polymorphisms (SNPs) associated with CVD, and used 0, 1 or 2 as the count of minor alleles for each SNP.

Deep learning models: To enlarge the statistic power, we pretrained a 1-D CNN on EHR dataset (removing the testing test) to learn representations of longitudinal EHR (Figure 1). We concatenated the pretrained CNN with an input layer with 204 SNPs and linked to a deep neural network with one hidden layer (dense_1). We used the genotyped individuals to train and validate the concatenated model.

Results
We identified 109,490 individuals, including 9,824 cases and 99,666 controls (mean [SD] age 47.4 [14.7] years; 64.5% female and 86.3% European) that met our inclusion criteria. Of them, 10,162 individuals were genotyped. We performed 10-fold cross-validations. Results showed the Area under the Receiver Operating Curve (AUROC) that using only EHR features is 0.704 (+/- 0.008), and the proposed model combing genetic features is 0.716 (+/- 0.006).

Conclusions
In this study, we proposed a concatenating framework of CNN to fuse heterogeneous features. Adding genetic features offered small but significant benefits to clinical features, which underline the importance of genetic data.

References
Summarization of Patient Education Material for New Media Platform

Yiqing Zhao, PhD¹, Yanshan Wang, PhD¹, Paul R. Kingsbury, PhD¹, Michael A. Panzer², Richard T. Van Ert², Hongfang Liu, PhD¹

¹ Division of Digital Health Sciences, Department of Health Sciences Research, Mayo Clinic, Rochester, MN.
² Division of Content Management and Delivery, Department of Global Business Solutions, Mayo Clinic, Rochester, MN.

Background
Printed or Internet-based patient education materials frequently consist of detailed information of disease conditions and procedures. However, they were usually designed for a general audience and sometimes fail to address the unique needs, interests, and concerns of different individuals. With the advent and dissemination of as interact new media platforms such as Amazon Alexa, our ability to understand audiences’ information needs has significantly increased. Also, considering the over-estimated health literacy among general population¹, we need to provide more brief and relevant information while still keep easy-to-understand explanation of some medical terms.

Methods
In this study, we used TextRank² to algorithm to summarize patient education documents. We tested on a subset of patient education documents focused on diseases and procedures. We first identified sections within each document and apply the algorithm on each section separately. We applied summarization ratio of 0.7 (percentage kept) on each section. We compared a sample our summarization results (20 documents on diseases and procedures respectively) against manually summarized documents. Summarization performance was evaluated using ROUGE-N metric³.

Results
We collected 3,286 patient education documents from Mayo Clinic Department of Global Business Solutions, which is also accessible from Mayo Clinic homepage (https://www.mayo Clinic.org). Among which 876 documents are Questions and Answers, 862 are Health Articles, 1,109 are about Disease Conditions, 22 are about First Aid tips, 88 are about Symptoms, and 329 are about Test Procedures. Sections for Disease Conditions include: Overview, Symptoms, Causes, Complications, Diagnosis, Treatment, Lifestyle and home remedies, Alternative medicine, Preparing for an appointment, Risk factors, Coping and support, and Prevention. Sections for Test Procedures include: Overview, Risks, Why it's done, What you can expect, How you prepare, and Results. Evaluation result is listed in Table 1. For Disease Condition documents, Diagnosis, Causes and Overview sections have relatively high F-score while for Test Procedures documents, Overview section has significantly higher F-score than other sections. Differences in performance might be due to difference in information needs for each sections.

<table>
<thead>
<tr>
<th>Disease Condition - Section</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
<th>Test Procedures - Section</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>0.553</td>
<td>0.464</td>
<td>0.497</td>
<td>Overview</td>
<td>0.631</td>
<td>0.745</td>
<td>0.674</td>
</tr>
<tr>
<td>Symptoms</td>
<td>0.436</td>
<td>0.293</td>
<td>0.326</td>
<td>Why it's done</td>
<td>0.500</td>
<td>0.255</td>
<td>0.284</td>
</tr>
<tr>
<td>Causes</td>
<td>0.696</td>
<td>0.485</td>
<td>0.538</td>
<td>Risks</td>
<td>0.396</td>
<td>0.743</td>
<td>0.441</td>
</tr>
<tr>
<td>Risk factors</td>
<td>0.390</td>
<td>0.390</td>
<td>0.390</td>
<td>How you prepare</td>
<td>0.417</td>
<td>0.821</td>
<td>0.463</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>0.533</td>
<td>0.606</td>
<td>0.563</td>
<td>What you can expect</td>
<td>0.400</td>
<td>0.313</td>
<td>0.341</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.408</td>
<td>0.408</td>
<td>0.408</td>
<td>Results</td>
<td>0.267</td>
<td>0.286</td>
<td>0.275</td>
</tr>
<tr>
<td>Average</td>
<td>0.503</td>
<td>0.441</td>
<td>0.454</td>
<td>Average</td>
<td>0.435</td>
<td>0.527</td>
<td>0.413</td>
</tr>
</tbody>
</table>

Conclusions
Our work demonstrates that using TextRank to summarize patient education material achieves good performance for some sections. From the results we can see that different sections have different performance. In the future work, we will focus on those sections with unsatisfactory performance and develop new summarization methods.

Reference
Automatically Identifying Intimate Partner Violence from Clinical Narratives for Emergency Department

Vivienne J Zhu, MD, MS,1 Christine Carr, MD,2 Kit Simpson, Ph.D,3 Roberta Dawson, RN, MBA,2 Alyssa Rheingold, Ph.D.,4 Leslie A Lenert, MD, MS1

1Biomedical Informatics Center, 2Department of Emergency Medicine, 3Department of Healthcare Leadership and Management, and 4Department of Psychiatry and Behavioral Sciences at Medical University of South Carolina, Charleston, South Carolina

Abstract: We developed and evaluated a natural language processing (NLP) approach for identifying intimate partner violence (IPV) from clinical narratives. 191,728 clinical notes from 12,276 adult female patients with emergency department visits were used as a training dataset to iteratively develop a lexicon of IPV, customized negations, and NLP algorithms. Through manual review of clinical notes for 312 NLP identified IPV patients (2.5%) in the training dataset, our NLP pipelines yielded 93.3% precision.

Background: Intimate partner violence (IPV) is a significant public health and criminal justice problem that negatively affects millions of victims yearly in the United States, primarily women (85%) of reported incidents. Although healthcare management associations recommend accurate documentation and coding of IPV incidents, accurate coding of IPV in medical records is still uncommon. In clinical practice, however, providers record clinical findings as narratives for each patient encounter, including symptoms, health histories, and results of physical exams. From these clinical narratives, NLP can identify and extract evidence of IPV-related suspected injuries and mentions of abuse.

Method: We used data from the Medical University of South Carolinas (MUSC) Research Data Warehouse. Female patients who were 18 years-of-age or older and had an Emergency Department (ED) visit at MUSC between 01-01-2017 and 12-31-2017 were eligible for this study. NLP pipelines identifying IPV were developed via extraction of clinical notes relevant to each patient’s ED visit. A de-identified subject-ID was used to link source documents and data across each patient’s records. Of 16,359 eligible patients, we randomly sampled 12,276 (75%) patients to create a training dataset. We used commercial NLP software (Linguamatics I2E version 5.3, Cambridge, United Kingdom) licensed to MUSC to index, parse, and query each research criterion. A set of NLP queries was used to capture semantic and syntactic representations of IPV. The following criteria were used to identify IPV mentions: 1) intimate partner violence (e.g., “domestic violence/abuse,” “intimate partner violence/abuse,” or ”abuse of partner”); 2) the combination of perpetrator (e.g., “husband,” “ex-husband,” “boyfriend,” “baby’s father,” “fiancé”) and abusive behavior (e.g., “hit,” “non-accidental injury,” “assault,” ”attack,” “threaten,” “verbally abuse,” “pull out hair”); and 3) social services (e.g., “My Sister’s House,” “Adult Protective Services,” “Verizon Hope”). Clinical negations were used to exclude false mentions of IPV, such as “no history of” and “denied”, or to exclude a family member as the IPV victim. Herein we report preliminary data for our NLP pipeline development and evaluation.

Results: Altogether, 191,728 clinical notes from 12,276 patients were used to develop the NLP pipelines. Without negation, 326 patients (2.6%) had IPV mention(s) detected from 948 clinical notes. With negation, 312 (2.5%) had IPV mention(s) detected from 907 clinical notes. The common type of providers who documented IPV incidents were physician, social worker, resident, registered nurse, and nurse practitioner. Among 307 terms (with morphologic variations) relevant to IPV, the most prevalent were “domestic violence,” “intimate partner violence,” “domestic abuse,” “Adult Protective Services,” and “assaulted by boyfriend”. Manual review identified 22 false positive IPV patients. The NLP pipeline demonstrated 93.3% precision in the training dataset. Example reasons of false positives included the term “Adult Protective Services” were suggested for other vulnerable patients (e.g., schizophrenia patient), as well as the following features: insufficient post negation for an IPV incident, and patients did not confirm providers’ suspicious or concern about IPV.

Conclusions: The results suggest that NLP can identify and extract evidence of intimate partner violence from clinical narratives obtained during and after ED visits, with high precision for female patients.

Acknowledgement: This research was supported in part by NIH grants (NLM R21LM012945, NIDA 2K12DA031794-06A1, and NCATS 5UL1TR000062-05) to the Medical University of South Carolina.
Bootstrapping a Mobility Dictionary from a Seed Set

Ayah Zirikly, PhD1, Denis Newman-Griffis, MS1,2, Julia Porcino, MS1, Pei-Shu Ho, PhD1, Jonathan Camacho Maldonado, MD1, Maryanne Sacco, MA1, Alex Marr, BA1
1National Institutes of Health Clinical Center, Bethesda, MD; 2The Ohio State University

Introduction

Limitations in ambulation and mobility are the top disability types that affect individuals1. In order to detect and monitor these limitations, it is essential that information about whole-person mobility is reported consistently. The International Classification of Functioning, Disability and Health (ICF), developed by the World Health Organization2, includes coding standards for mobility. However, low adoption of ICF coding in clinical notes makes automated identification of information about individuals’ function challenging and expensive, especially compared to the ready availability International Classification of Diseases (ICD) codes for diseases. New terminologies can ease the identification and extraction of mobility information. In this work, we expand a manually-extracted seed dictionary to capture terms that are linked to ICF concepts about mobility, such as walking, carrying, and transferring body positions.

Methods

We start with a dataset of 450 physical therapy notes (∼260k words) collected from the NIH Clinical Center (BTRIS), in which mobility concepts have been manually annotated with ICF codes. We extract a seed set containing the high-frequency mobility terms (keywords) from these data, and introduce two methods to expand this seed set. Our first expansion methodology uses VerbNet3, a lexicon that clusters verbs that show semantic coherence into classes. However, this method suffers from multiple limitations: the expansion only considers unigrams in isolation, and because all VerbNet class members have equal weight, they cannot be ranked and filtered as expansion terms. Therefore, we propose another expansion method that overcomes these limitations. We use word vectors trained from PubMed data with FastText4, and sum-vector for multi-words terms. Then, we retrieve the ten nearest neighbors (NN) to the original keyword, based on the cosine similarity between the vectors for original keyword and a candidate term. This allows words to be compared on their semantic properties in context rather than form alone.

Results

We test the recall of the expanded dictionary resulting from the two approaches using exact and partial match metrics. Partial match evaluation is done at the token level, where an overlap of one token between the human-annotated mobility term and the dictionary entry will yield a match. In exact match, all the tokens need to be identical. Our combined dictionary yields a recall of 60.35% for exact match, where the VerbNet-based approach has a recall of 16.39%, and the semantically-aware approach yields 50.7%. The partial match results are significantly higher with recall 95.66% for the combined dictionary, in contrast to 32.07% and 87.96% for VerbNet and NN approaches respectively. Results analysis in the exact match evaluation indicates that non-matches are mainly due to: i) Specific punctuation sequences that are hard to match, for instance bed to w/c where w/c is wheelchair; ii) Prepositions in the multi-term expressions (e.g. sit to stand vs. sit-stand). These cases convey the need to improve preprocessing on the seed set and postprocessing on the expansion terms. In addition, a “new” exact-match evaluation can be introduced that does not require full overlap, but rather ignores cases of punctuations and prepositions.

Conclusion & Future Work

We find that a combination of lexicon- and context-based expansion methodologies yield high recall on mobility terms, and a nearest neighbor approach based on word embeddings significantly outperforms lexicon-based expansion. In current work, the expansion terms inherit the ICF code of the original term. In future work, we will expand to methods that infer the correct ICF code from all the other semantically-close terms in the seed set. Additionally, we will investigate aligning our mobility dictionary with SNOMED concepts for further expansion. We will also develop post-processing functions to handle misspellings that are not handled by our expansion methodologies.

References

Authoring and Integrating Interoperable Clinical Decision Support: CDS Connect Open Source Tools

Shafa Al-Showk, MPH¹, Chris Moesel², Sharon Sebastian, MS, RN-BC², Ginny Meadows, MSHI, RN-BC², Steven Bernstein¹, Mary Nix, MS, PMP¹, Edwin A. Lomotan, MD¹
¹Agency for Healthcare Research and Quality, Rockville, MD; ²The MITRE Corporation, Bedford, MA

Background

In its 2017 special publication, Optimizing Strategies for Clinical Decision Support, the National Academy of Medicine identified several barriers to realizing the full benefits of clinical decision support (CDS), including the high cost of developing reliable, shareable CDS expressions and the absence of systematic means to validate CDS content. The Agency for Healthcare Research and Quality (AHRQ) has been creating prototype infrastructure to ease the authoring and testing of interoperable, standards-based CDS expressions that can be ultimately shared through a national CDS repository called CDS Connect. The aims of CDS Connect, together with other ongoing AHRQ CDS initiatives, are to advance evidence into practice through CDS and to make CDS more shareable, standards-based, and publicly-available.

Description of Demonstration

AHRQ, through the CDS Connect project, has been constructing open source tools to reduce the burden of developing accurate, trustworthy interoperable CDS expressions that can be integrated with health information technology (IT) systems to support evidence-based decision making. The tools include 1) the CDS authoring tool, which provides a user-friendly interface that lowers the barrier to creating standards-based CDS expressions and 2) Clinical Quality Language (CQL) Services, which facilitates integration of the CDS developed using the authoring tool with a health IT system.

The CDS authoring tool empowers non-software engineers to design interoperable CDS logic according to the CQL standard. A key feature of the authoring tool is its ability to interact with the National Library of Medicine’s Value Set Authority Center (VSAC). The CDS authoring tool allows users to build CQL by stringing together value sets, operators, and other connectors to define the logic that can drive CDS. It also supports user-testing of authored artifacts using test data that can be created manually or by exporting synthetic patients using the Synthea project. In addition, users can download their tested CQL logic using the HL7 FHIR DSTU2 model or the STU3 model.

CQL Services is prototype software that allows CQL execution to be externalized via RESTful web services. A health IT system can invoke CQL-based CDS logic by sending web requests to CQL Services, which returns the calculated results back in JavaScript Object Notation (JSON) format. The health IT system can then display the results via their native user interface. CQL Services exposes two application programming interfaces (APIs); a simple custom API for exchanging data and results using the JSON format, and a standards-based API that conforms to the CDS Hooks standard for interacting with external CDS services.

In 2019, AHRQ and its partners, b.well Connected Health and the MITRE Corporation, are developing and testing CDS to support preventive health care. As a use case, this demonstration will highlight how the CDS authoring tool can be used to create and test CDS such as this, and how CQL Services can be used to integrate the validated CDS in a health IT system for clinical testing and implementation.

During this presentation we plan to demonstrate: How to build CQL-based clinical decision support artifacts using the CDS Connect Authoring Tool; how authors can incorporate value sets from the National Library of Medicine’s Value Set Authority Center into their own CQL-based clinical decision support; how the developed CQL-based clinical decision support can be tested against a synthetic dataset; how the CQL Services prototype tool can integrate CDS logic with a health IT system using CDS Hooks standards.

Author Statement

CDS Connect was launched in spring 2017. The CDS authoring tool and CQL Services prototype tools are freely available, open source, and accessible at https://cds.ahrq.gov/cdsconnect/authoring and https://github.com/AHRQ-CDS/AHRQ-CDS-Connect-CQL-SERVICES. Information about the AHRQ initiative, including access to CDS Connect, can be found at http://cds.ahrq.gov.
Background. The National Committee for Quality Assurance (NCQA) recently announced Population Health Management (PHM) as a new standards category to shift focus from single disease to complex case management and wellness by promoting strategies that help coordinate patient care between providers and practices more efficiently. The new standards support and emphasize using data analytics to target resources to the right individuals, but methods for standardizing data analytics and population-based patient care workflows have not yet been specified. Most market-leading EHR vendors provide PHM tools, but solutions are proprietary, difficult to share, and therefore not optimally scalable. Standards-based PHM methods and tools are needed to scale beyond single vendor solutions and to reduce the implementation barriers related to reproducing complex case logic and analytics.

Through a research grant funded by the Informatics Technology for Cancer Research (ITCR) program of the US National Cancer Institute (NCI), we have devised and implemented a PHM platform that adapts a synchronous and individual-level clinical decision support (CDS) standard (the HL7 CDS Hooks standard in conjunction with HL7 FHIR) to an asynchronous population-level CDS via Web services. The primary methods added to the existing standards to support population-level CDS were bulk operations. Rather than creating one CDS Hooks request at-a-time for each patient, CDS Hooks requests were composed in bulk from population-based queries. Bulk CDS Hooks requests and responses were then composed, sent and received by the CDS engine in large batches to reduce network request/response times. Preliminary performance measures have demonstrated these simple batch operations reduce the processing time required by two orders of magnitude.

Use Case. We will demonstrate an application of the PHM platform, which screens and identifies patients who meet National Comprehensive Cancer Network (NCCN) criteria for genetic evaluation of familial cancer risk based on their family history in the EHR. The application uses both structured data and natural language processing of free-text data. Patients identified by the standards-based PHM platform are imported into an EHR's PHM dashboard (e.g., Epic's Healthy Planet) where genetic counseling staff can review individual cases, select, and send bulk outreach messages to patients suggesting they setup genetic counseling appointments. Additionally, the PHM communicates care decisions with the patient's primary care provider to ensure care is coordinated.

A pilot implementation of the PHM platform has been live in the University of Utah Health system and Genetic Counseling Service at the Huntsman Cancer Institute since October of 2018. An NCI-funded, multi-site randomized controlled trial testing two alternate patient outreach approaches is underway. The CDS Hooks services that implement the NCCN familial cancer risk detection algorithms are available to the community as open source software.

The speakers, who are national leaders in standards-based CDS innovations, will describe (i) the PHM workflow; (ii) the family history-based criteria and algorithm logic; (iii) performance optimizations that were made to make population analysis using CDS Hooks viable; (iv) preliminary results of the pilot study; (v) ongoing efforts to integrate the PHM platform with Epic at New York University (NYU) and Cerner at Intermountain Healthcare; and (vi) a discussion of challenges and opportunities for future research leveraging the PHM platform.

Acknowledgement. This project is supported by grant U24CA204800-01 from the Informatics Technology for Cancer Research (ITCR) program of the National Cancer Institute.
ROAD2H: Learning Decision Support System for Low- and Middle-Income Countries

Kristijonas Čyras, PhD¹, Jesus Dominguez, PhD², Amin Karamlou, MSc¹, Denys Prociuk, MSc¹, Vasa Curcin, PhD², Brendan Delaney, MD¹, Francesca Toni, PhD¹, Kalipso Chalkidou, MD PHD¹, Ara Darzi, MD¹
¹Imperial College London, London, UK, ²King’s College London, London, UK

Description

Learning Health Systems offer great potential for Low- and Middle-Income Countries to develop better application of evidence-based practice, whilst at the same time being able to learn about and adapt, rather than simply import evidence and guidelines from more advanced economies. Opportunity is provided for both patient-facing decision support and health-system facing views, the latter being particularly important for enabling planning and resource allocation. In addition, decision support systems should enable transparency of decision making to users and allow for support in prioritization and resolving conflicts between guidelines and individual patient status. ROAD2H (Resource Optimisation, Argumentation, Decision support and Knowledge Transfer to create Value via Learning Health Systems) is a UK EPSRC Global Health Project working in partnership with China and Serbia to provide both individual patient decision support for clinicians and a health system dashboard in the following areas:

- In Belgrade, Serbia, supporting both clinical pathways for the admission and treatment of Chronic Obstructive Pulmonary Disease (COPD) patients with acute exacerbations, and transition of care for community-based patients with COPD between specialist and primary care.
- In W. China (Wushan and Ninquiang), supporting transition of care for COPD between township clinics and county hospitals.
- In Beijing, supporting efficient use of high-cost drugs for Non Small Cell Lung Cancer (NSCLC).

The system takes a combination of local, regional, national and international guidelines, represents them using a logical model (Transition-based Medical Recommendations, TMR¹) and adds an additional argumentation-based reasoning layer² to aid in resolving conflicts among patient-tailored guideline recommendations, while taking into account multiple goals with differing priorities and respecting the patient’s preferences³. Cost effectiveness and willingness to pay for proposed interventions are optimized using a stochastic discrete-event simulation model programmed in R, parameterised using both local clinical and insurance data, and incorporated as an ‘expected’ outcome for health policy. Data is gathered via the EHR system locally and used to drive a dashboard, and potential changes to policy that result in changes to individual clinical recommendations.

Having worked extensively with users on use-cases and data modeling, we have developed an end-to-end exemplar in community-based COPD in the TMR combined with argumentation. We aim to demonstrate a use-case with patient EHR passed through a decision support system operating with interacting COPD guideline recommendations (via TMR) and resolving their interactions (via argumentation) in accordance to the patient’s context, their preferences and/or priorities over goals.

References

Towards a digital platform for management of randomised controlled trials across multiple electronic health record vendors

Jesús Domínguez, PhD, Elliot Fairweather, PhD, Brendan C. Delaney, MD, Vasa Curcin, PhD

1King’s College London, London, UK; 2Imperial College, London, London, UK

Introduction

Although randomised controlled trials (RCTs) are the most reliable approach to produce high quality medical evidence, protocols for RCTs are slow to produce and often imprecisely specified, resulting in inefficient recruitment and execution. This situation is exacerbated when managing multiple configurations across clinical sites, as required to run trials in a primary care setting. We offer a digital clinical trial platform, or study system (SS), which integrates clinical research and patient care workflows through Electronic Health Records (EHR) systems, taking advantage of initiatives such as CDISC, which provides standards for computational clinical trial definitions, and advances made in data exchange between EHRs and electronic Case Report Form (eCRF) components enabling semantic inter-operability across vendors and automated pre-population of eCRFs from EHR data.

System Design and Features

First, the study protocol is formalised following an extension of CDISC operational data model (ODM) where we have incorporated EHR queries by means of openEHR archetypes for the identification of particular clinical elements (e.g., diagnosis tags, clinical codes, demographics, medication) used as flags to capture potentially eligible candidates. Other extensions include identifiers for retrieval of EHR-neutral, patient-specific data for eCRF pre-population, and attributes to convert item definitions into HTML5 form elements when requested by data node connectors (DNC) deployed at medical practices. The RCT database (DB) is then populated with CDISC representations of the study’s life cycle and eCRFs extracted from the formalised protocol. This is then used by the SS to coordinate study execution across distributed medical sites. Additionally, data-filled eCRFs are also stored in the DB, preserving patient anonymity by means of encryption. eCRFs are grouped by type and the SS acts upon them depending on the specifics of the trial, e.g., a participant’s email address is retrieved from any form of type Contact Details. Similarly, participant randomisation is supported by retrieving randomisation arguments from a submitted form and returning the outcome to the clinician via another form. A third (internal) form provides relevant parameters for the randomisation algorithm to be trial-specific, additionally generating a username and password for participants to access the dedicated web app. Identification and notification of trial-specific events such as form completion reminders are also managed by the SS using text messaging and/or email address. The web app is designed to support patient data collection via form templates. It is accessible to both participants and study members in a study-specific manner with login details stored in an LDAP server. The DNC mediates the relationship between a vendor’s EHR and the centralised SS, adapting to distinct EHRs by means of a vendor-specific plug-in and downloading trial-specific forms when the timing is appropriate. Downloaded eCRFs are pre-populated with patient-specific data and integrated in the EHR when submitted. An embedded Chromium web browser displays eCRFs independent of local browser configuration, providing HTML5- and CSS3-compliant elements while taking advantage of HTML5 constraint attributes and types to validate data.

Current Use and Deployment

The SS is currently being deployed at clinical sites across UK to support a two-year study on whether Ciprofloxacin drops or delayed oral Amoxicillin are non-inferior to current standard practice for children with acute otitis media with discharge (https://doi.org/10.1186/ISRCTN12873692).

Conclusion

We described modeling aspects of the implementation of a study system that addresses the integration of clinical trials into routine clinical practice, reducing both the cost and complexity of the task whilst increasing accuracy and safety by liaising directly with the EHR system. This is achieved by making a functional eCRF semantically interoperable with the EHR system in order to enable patient identification and pre-population of eCRFs with available clinical data.
Using trial2rev to support timely and efficient systematic review updates

Adam G. Dunn, PhD1,2, Paige Martin, BE1, Florence T. Bourgeois, MD MPH2,3
1Centre for Health Informatics, Macquarie University, Sydney, NSW, Australia;
2Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA;
3Departments of Pediatrics & Emergency Medicine, Harvard Medical School, Boston, MA

Abstract
Systematic reviews are time-consuming and often do not target the most relevant medical questions, leading to slow uptake of new trial results and many potentially redundant reviews. We developed the web-based platform trial2rev to connect published systematic reviews to emerging results in clinical trial registries, with the aim of supporting decisions around when a systematic review update should be undertaken. Major use cases include the proactive monitoring of emerging trial results to signal when a review update is warranted and using the database of systematic reviews linked to trial registrations to support evaluation of new methods for automatic screening of trials.

Demonstration
Even for a well-studied area like type 2 diabetes, where as many as 23 new systematic reviews on drug interventions are published each year, it takes a median of 1.5 years for the results of a newly-published drug trial to be included in a systematic review.1 We developed the trial2rev platform in 2018 to help systematic reviewers monitor the accumulation of relevant trial data as a way of quickly determining when a systematic review update is warranted.2 Trial2rev (https://surveillance-chi.mq.edu.au) is a web-based platform and database maintaining structured information on trial-based systematic reviews. It links reviews to the registrations and publications of trials included in the review and provides lists of other similar systematic reviews with overlapping sets of trials. Trial2rev uses PubMed, CrossRef, and ClinicalTrials.gov to identify new systematic reviews and trials. Multiple software agents using machine learning methods are triggered to recommend trials,3 as users interact with the system. While machine learning methods are used to help screen published articles,4 this functionality has not yet been extended to support the use of trial registrations. The trial2rev system enables proactive surveillance of clinical trial registrations to support decisions about when to update a systematic review, a paradigm shift from the traditional workflow of deciding to undertake a systematic review and then reactively screening for relevant trial publications. Trial2rev has not yet been evaluated against manual processes that are currently used to decide when to update a review.

The first use case for the platform is to support systematic reviewers deciding if a systematic review update is warranted. Users start by identifying a systematic review, verify information about the trials that were included, and examine the registrations of ongoing and completed trials that have been recommended as relevant by software agents or human users. Users can flag reviews of interest to receive alerts about the statuses of relevant trials.

The second use case supports the clinical research informatics community with the data and infrastructure needed to test and deploy methods for automated screening of trials for inclusion in systematic review updates. Registered users gain access to the large database of linked systematic reviews and trial registrations to train and test new methods for identifying trials for systematic reviews; and may deploy new methods and software agents in trial2rev.

The source code for trials2rev is available on GitHub and its functionality is updated periodically. The platform currently includes partial structured information for 13,477 systematic reviews and 14,733 linked trials, and new data are made available daily. The research was supported by the Agency for Healthcare Research & Quality (R03HS024798) and the National Library of Medicine (1R01LM012976-01).

References
Demonstrating the Collection and Integration of Standardized Patient-reported Outcome Data Using a SMART on FHIR® Application

Chun-Ju Hsiao, PhD1, Joseph Blumenthal, BA2, Deliya Wesley, PhD3

1Agency for Healthcare Research and Quality, Rockville, MD; 2MedStar Health National Center for Human Factors in Healthcare, Washington, DC; 3MedStar Health Research Institute, Hyattsville, MD

Background

The effective use of patient-reported outcomes (PROs) can play a critical role in improving health care delivery and patient experience with care. While various efforts promote the use of PROs, their uptake is far from universal. Some electronic health records (EHRs) are able to capture structured PRO data, but this information is not commonly collected at the point of care. In addition, standards have not previously existed to guide and inform the collection and integration of PRO data into health information technology (IT) systems, thereby limiting the ability to easily share these data across health systems. To fill this gap, the Agency for Healthcare Research and Quality (AHRQ) is working with partners to support the development and implementation of electronic applications that can enable the collection and sharing of standardized PRO data across health providers.

Description of Demonstration

In 2018, the Office of the National Coordinator for Health Information Technology created a Health Level 7 International (HL7®) Fast Healthcare Interoperability Resources (FHIR®) PRO Implementation Guide (IG) to address the previously mentioned challenge regarding lack of standards. Together with its partners (MedStar Health and Outcomes Based Electronic Research Database (OBERD)), AHRQ modified an existing mobile application that incorporated the FHIR® PRO IG to enable the collection and integration of standardized PRO physical function data. The modified application can administer Computer Adaptive Tests (CAT) to collect data on patient’s physical function using the state of the art PROMIS® measures. The modified application has been used with three different EHR systems, and demonstrates the interoperable, scalable, and replicable nature of this solution. MedStar leveraged the same technology and processes as those used in most EHR’s data views to surface the PRO data collected from the application. Thus, the PRO data can be included as part of a release-of-information request in the EHR so that PRO information can be used in future medical decision-making.

The architecture of the PRO data collection and visualization ecosystem features a decentralized, scalable, and replicable model. The architecture enables EHR agnostic data ingestion, patient matching, and visualization of these data. This demonstration will include a live walkthrough of the data flow as it begins with the patient accessing a mobile application that administers the CAT by connecting with a SMART on FHIR® external assessment center. The audience will observe how the application is used to collect PROMIS® measures, connects to a secure FHIR® endpoint, and sends these data for storage and subsequent retrieval. The demonstration will also include a customized data visualization of the PROMIS® scores and survey responses that healthcare providers would see in the EHR. In addition, the demonstration will show the “data on demand” feature that expedites site-specific implementation by eliminating the need to create data feeds through proprietary interface engineering.

This demonstration will highlight the unique architecture and how it leverages the power of SMART on FHIR® and the FHIR® PRO IG to create an extensible, light-weight, and portable solution.

Author Statement

The PRO project was launched in Spring 2018. The modified application pilot test plan was finalized in Fall 2018 and the application redesign took place in Winter 2018. Application implementation at nine primary and specialty care practices began in February 2019 and is expected to be completed by June 2019. Evaluation of application usage and users’ feedback will occur in Summer 2019. All authors listed above contributed to the application’s development and implementation. Chun-Ju Hsiao and Joseph Blumenthal will conduct the demonstration.
Evidence-Based Care Made Easy: University of Utah’s SMART on FHIR Platform for Chronic Disease Management and Health Maintenance

Kensaku Kawamoto, MD, PhD, MHS1, Douglas K. Martin, MD,1 Claude Nanjo, MAAS, MPH1

1Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Problems addressed: crises in care quality and clinician burnout. The U.S. faces a care quality crisis, with U.S. adults receiving only about 55% of recommended care.1 The U.S. also faces a crisis in clinician burnout, and a significant reason is that for every hour of clinical work, clinicians spend 2 hours on clerical work or EHR-related tasks in the ambulatory setting.2 These two issues are likely related, as EHRs generally provide only limited support for delivering evidence-based care efficiently and effectively, contributing to both clinician frustration and suboptimal compliance with evidence-based care standards.

University of Utah ReImagine EHR initiative. In 2016, the University of Utah launched the ReImagine EHR initiative to improve patient care and the provider experience through standards-based, interoperable extensions to the EHR. Supported through operational investments and over $27 million in grants and contracts, this initiative has developed numerous informatics solutions that leverage the HL7 SMART on FHIR and CDS Hooks interoperability standards. Systems developed and deployed into production in the health system’s Epic EHR environment include a neonatal bilirubin management application, a population health management system for individuals at risk of early onset familial cancers, a surgical referral dashboard, a procedure capacity management application, a lung cancer screening shared decision making app, a diabetes treatment outcome prediction app developed in collaboration with Hitachi, opioid-related decision support tools developed with CDC and ONC support, and EHR-integrated medical calculators developed in collaboration with MDCalc. Several of these solutions have been integrated with multiple EHR platforms. The ReImagine EHR team includes seven members certified to develop new FHIR interfaces into the Epic EHR. One of these apps (the bilirubin management app) has won several provider user experience awards.4

SMART on FHIR Disease Management Platform. Building on the expertise gained from the efforts above, the ReImagine EHR initiative identified evidence-based chronic disease management and health maintenance as a key need for improving care quality and the provider EHR experience. To address this need, the Disease Management Platform was developed. This system is fully integrated with the EHR, and it provides care guidance via individual chronic disease modules as well as a unified module that encompasses all relevant modules for chronic diseases and health maintenance. The initial disease module supported by the platform is chronic obstructive pulmonary disease (COPD), and additional modules are in process such as for hypertension, diabetes, general health maintenance, congestive heart failure, and chronic kidney disease. The system is available in the production Epic environment. In order to maximize potential impact, a commercial approach to disseminating the system is being explored.

Key Innovative Features. As one key innovative feature, the Disease Management Platform integrates advanced, guideline-based decision support with commercial EHR technology. Second, the system is able to utilize practically any data available in the EHR, regardless of whether it is currently supported by the base FHIR interface provided by the EHR vendor. This is made possible by additional FHIR interfaces developed by the ReImagine EHR team, and which can be deployed at other healthcare systems. Third, the system pulls in and intuitively organizes relevant data, so that there is no need to hunt for the relevant data across the various modules in the EHR. Fourth, the system provides a unified summary of the patient’s status on guideline-based care standards, as well as actionable recommendations on next steps. Fifth, recommendations can be followed easily, such as through 1-click ordering of needed interventions. Finally, the system facilitates required clinical documentation.

Session Content and Demonstration. The presenters will start by briefly describing the problems addressed, the University of Utah ReImagine EHR initiative, and the approach taken to design and implement the Disease Management Platform. The presenters will then use sample patients to demonstrate how the Disease Management Platform makes it easy – and perhaps even enjoyable – to provide evidence-based care to every patient, every time. The presenters will conclude with a brief summary of evaluations to date, lessons learned, and future directions.

A Molecular Research and Discovery Platform for Placental Research: the Placental Atlas Tool (PAT)

Michael Keller, Ph.D.¹, John Ilekis, Ph.D.², Alexandra Shlionskaya, M.S.¹, Christopher H. Ferguson, Ph.D.¹, Archana Mohale, M.S.²

¹Booz Allen Hamilton, Inc. McLean, Virginia, USA; ²Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland, USA

Introduction

The Placental Atlas Tool (PAT) (https://pat.nichd.nih.gov/) has been developed to create, maintain, and update a shared research analytics platform designed to serve as an important resource for the placental research community. PAT leverages existing resources to identify knowledge gaps, develop new hypotheses to address those gaps, and identify potential therapeutic targets and biomarkers. PAT addresses a need for scalable and agile systems integration, supporting NICHD’s organizational goals and their user community, and maximizing research investment through data discoverability and reuse.

PAT Technical Implementation

The PAT system employs a microservice architecture methodology to be scalable, extensible, and data-driven to adapt to changing scientific and business needs of the placental research community. PAT was designed to be metadata-driven and built around standardized terminologies of concepts relevant to placental research which were used to identify and annotate all datasets, images, and publications available in the system. Molecular datasets are preprocessed and loaded into the PAT database to be easily discovered and analyzed by users with embedded web-based analytical tools.

PAT User Experience

PAT is based on user-centric design principles and places emphasis on a flexible, data-driven experience that enables discovery from multiple user perspectives. The PAT Homepage provides access to the various explorers as well as education and information resources. The Terminology Browser contains the terminologies that form the basis of the metadata-driven content experience. The Image Explorer allows users to search and sort through placental images enriched with metadata. The Dataset Explorer provides users with access to over 500 placental datasets. The dataset analytics workflow makes bioinformatic analysis of large molecular datasets accessible to all users, regardless of bioinformatic experience and provides common bioinformatic outputs. Users are able to save content entities of interest to their personalized Workspace. Figure 1 provides snapshots of the PAT functionalities listed above.

Figure 1. PAT Functionalities

Conclusion

PAT is poised to maximize the investment NICHD and the broader research community has made in placental research by reinforcing open scientific inquiry, expediting formation of new insights by combining data sources, promoting the testing of new or alternative hypotheses and analytic methods, and facilitating education of new researchers. PAT accomplishes this by coupling cutting-edge data-driven, flexible development principles with rigorous content curation practices, ultimately creating a fluid, accessible, and robust investigational resource.
NIAID Tuberculosis Data Exploration Portal: A Public Analytics Platform

Alyssa R. Long1, Alexander J. Glogowski, MBA1

1Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA

Introduction

Tuberculosis (TB) is the leading cause of death from infectious disease worldwide, causing over 1.5 million deaths in 20171. Drug-resistant forms of TB present a rising public health crisis, because they are difficult to treat with lengthy and more expensive multi-drug regimens. Diagnosis and treatment for drug-resistant TB often requires a full clinical picture with radiological and pathogen genomic information, yet there was a lack of public data repositories that contained this information as a resource. Responding to this need, the National Institute of Allergy and Infectious Diseases (NIAID) started to collaborate with physicians, researchers, radiologists, and IT specialists in countries with a high burden of drug-resistant TB, forming the framework of the TB Portals Program2. Collecting, validating, analyzing, and annotating a wide range of TB data, we built an open-access repository and developed a user-friendly and powerful advanced analytics resource, TB Data Exploration Portal (TB DEPOT). TB DEPOT facilitates the visualization and analysis of multidimensional TB Portals Program data, which enables researchers to move from data to knowledge to action in the fight against TB.

Tuberculosis Data Exploration Portal

TB DEPOT integrates deidentified clinical, socioeconomic, genomic, and imaging data and allows for user-driven, repeatable, and reproducible statistical analyses. It provides a web-based analytics platform that makes unique, inaccessible drug-resistant TB data available to the research community, and enables the reusability of data, in conformity with the NIH’s Findable, Accessible, Interoperable, and Reusable (FAIR) principles. TB DEPOT provides a self-service capability for users to create, compare, and analyze patient cohorts to generate and test their hypotheses, going from thought to analysis in under a minute. The cloud-based tool is hosted on AWS, designed with an extensible, scalable infrastructure using .NET webforms, Highcharts, R, PLINK, custom algorithms, and provides APIs for data and cohort sharing, all of which can be replicated for other infectious or chronic diseases.

Figure 1.

TB DEPOT, deployed in Production in 2016 and publicly available at: https://depot.tbportals.niaid.nih.gov.

Conclusion

TB DEPOT is the first tool of its kind to facilitate open-access cohort creation, comparison, and analysis through the integration of clinical, socioeconomic, genomic, and imaging data on a web-based platform in a reusable way.

References

Interoperable Consumer Decision Support: CDS Connect and b.well

Chris Moesel1, Kristen Valdes2, Ginny Meadows, MSHI, RN-BC1, Shafa Al-Showk, MPH3, Quyen Ngo-Metzger, MD, MPH3, Sharon Sebastian, RN-BC, MS1, Edwin Lomotan, MD3

1The MITRE Corporation, Bedford, MA; 2 b.well Connected Health, Inc., Baltimore, MD; 3Agency for Healthcare Research and Quality, Rockville, MD

Background
Since 2016, the Agency for Healthcare Research and Quality (AHRQ) has been developing prototype infrastructure for developing and sharing interoperable clinical decision support (CDS). The infrastructure, called CDS Connect, includes a repository of publicly-available CDS knowledge artifacts and open source software. To demonstrate use of the infrastructure and tools, CDS Connect develops, implements, and shares CDS through use cases in specific clinical domains. In 2019, CDS Connect, in collaboration with the MITRE Corporation and b.well, developed interoperable decision support focused on preventive care. Recommendations were delivered directly to consumers through a comprehensive, patient-facing mobile platform.

Description of Demonstration
The U.S. Preventive Services Task Force (USPSTF) is an independent, volunteer panel of national experts that works to improve the health of all Americans by making evidence-based recommendations about clinical preventive services. The USPSTF works with stakeholders in the delivery of primary care to help communicate and implement its recommendations, which span a range of topics. CDS Connect worked with b.well to transform a subset of USPSTF grade A and B recommendations to CDS expressions for delivery through b.well’s consumer-facing platform.

b.well is a consumer health management platform that connects consumers with a single digital experience to manage all aspects of their health. The platform allows consumers to access their own medical records electronically, receive personalized services and recommendations based on their health needs, view insurance and financial information, and be incentivized to engage in their own health journey. The platform pulls data from electronic health records, health insurers, prescriptions, lab systems, health devices (e.g., fitness trackers), and more while considering patient preferences. The b.well platform organizes a consumer’s experience as a health journey in which they can play games that encourage healthy choices and behaviors, earn rewards for meeting personalized health goals, and receive reminders and incentives for preventive screenings and other ways to manage health for themselves and their family.

The CDS for USPSTF recommendations was developed by translating recommendation statements into fully specified, standards-based representations. Logic statements were encoded in HL7 Clinical Quality Language (CQL) and include value sets – groupings of codes from standard terminologies such as SNOMED-CT, LOINC, and RxNorm – to define clinical concepts. The CDS data model was defined in terms of the HL7 Fast Healthcare Interoperability Resources (FHIR) Draft Standard for Trial Use (DSTU) 2 standard for health data exchange. CQL Services, an open source prototype tool developed by CDS Connect, exposed the CQL-based CDS over a HL7 CDS Hooks API, allowing b.well to invoke the CDS to receive appropriate USPSTF recommendations for each consumer.

The CDS presented through b.well is consumer specific, i.e., tailored to their personal health information and health history. The platform recommends appropriate screenings identified through data such as age, gender, various risk factors, insurance claims, and other aggregated data. Consumers are notified that they are due for a screening via their preferred communication method. The b.well platform uses gamification and behavioral science to educate consumers about the screening and to provide appropriate actions, such as scheduling an appointment with their doctor. Consumers earn points for each task completed through the platform and can then use those points to redeem healthy rewards. Attendees will experience this health journey from the perspective of a consumer presented with USPSTF recommendations and will learn how to access the publicly-available, interoperable CDS artifacts on CDS Connect.

Author Statement
CDS Connect has been operational and widely available since 2017. USPSTF recommendations will be pilot-tested and implemented via the b.well platform in spring/summer 2019. All CDS knowledge artifacts, including the CQL code, results of the pilot, and a detailed implementation guide, will be publicly-available on CDS Connect by September 2019. Information about the AHRQ initiative, including access to CDS Connect, can be found at http://cds.ahrq.gov. All authors listed above contributed to the system’s development. Chris Moesel and Kristen Valdes will conduct the demonstration.
MSK-EXTRACT’s Shopping Cart: Integrating Ontology with REDCap

John Philip¹, MS, Rimma Belenkaya¹, MS, and Peter Stetson¹, MD, MA
¹Memorial Sloan Kettering Cancer Center, New York, NY

Introduction

Memorial Sloan Kettering uses an EHR with clinical notes that is comprised of free text. Despite the advances in natural language processing and other types of machine learning, much of structured clinical data at MSK is a result of manual abstraction, often stored in spreadsheets. This approach is problematic since spreadsheets are often lost, don’t use standard terminology and are not used in a multi-user fashion. REDCap is a PHI friendly, multi-user alternative to spreadsheets. Many institutions try to make REDCap more approachable by teaching classes and using data analysts to help people build their projects. However there remains a steep learning curve in starting a data curation project in REDCap vs spreadsheet. At MSK, the MSK EXTRACT initiative is implementing a three-tier approach to lure people away from spreadsheets to REDCap: a “shopping cart” application to build the REDCap database, dynamic data pull from a data warehouse and an ETL script that sends the abstracted data back to the data warehouse for reuse and analytics. Here we present the “shopping cart” application that enables the use of standard terminology in REDCap databases.

Shopping Cart Application

The Shopping Cart is a web application that helps researchers build REDCap databases built on data elements that have been mapped to ontologies by our terminology team.

Features:

• Add and delete fields. Add and delete permissible values for the fields.
• Fields can be public or private. Public fields are available for Dynamic Data Pull but only the labels are editable. They can also be added to new projects.
• Suggest new fields for terminologists to vet and add to Standard Library of terms after being mapped to ConceptIDs. Such fields are always public.
• Help user organize fields into REDCap screens (aka instruments).
• Use REDCap APIs to update database with changes.
• Use of existing projects as templates to build ones own project.

Conclusion

The key innovations in the shopping cart are: 1. Allowing a crowdsourcing approach to building a standard library. 2. Standard term labels are still customizable unlike most other terminology approaches that lock down terms. 3. The conceptIDs are hidden from the users, keeping people oblivious to the mechanism of standardization while still getting the benefits. 4. Standard terms can be added from the standard library or from other REDCap projects.

Current Deployment Status

As of current writing, the software is in beta. Custom APIs were built to allow communication to TopBraid and to a custom database (EXTRACT DB) stored in DB2. The shopping cart uses the REDCap APIs to update the REDCap database. The application is written in Python, using REACT for front end and Tornado for web framework. In the next few months, we plan to add: 1. a caching database for the standard library to avoid constantly querying TopBraid, 2. chat functionality on a field and project level to allow easy communication between user and the terminologist, 3. Additional search function beyond text search to rapidly find the desired term in the standard library. Recent UI exploration suggests a navigation with a tree structure might work very well.

1842
Killing Cancer With Data: Solving Cancer Data “Interoperability” With Portability Through The Patient

Anil Sethi
Founder and CEO
Ciitizen Corporation
Palo Alto, California, U.S.A.

Reducing cancer-related morbidity and mortality requires greater access to data. But today, across the globe valuable health data is trapped in silos. Institutional data holders are often reluctant to share data, either due to regulatory concerns, competitive concerns (often masked as regulatory concerns), or lack of sufficient incentives or resources to share. It is difficult and costly to find patients to enroll in cancer clinical trials and even existing cancer data initiatives suffer from critical information gaps.

Cancer has a data problem; let patients help solve it.

Ciitizen is developing a patient-controlled cancer data ecosystem that will enable patients to fill these data gaps. Patients already have the right to all of their health information under the Health Insurance Portability and Accountability Act (HIPAA) and other privacy laws. Leveraging this right of access, on behalf of the patient Ciitizen collects comprehensive health information from every place where the patient has received care, from pre-cancer diagnosis forward, including images, notes, pathology reports, genomic/genetic data, and other data not yet available through patient portals. Ciitizen then normalizes and standardizes this data into actionable information (with data provenance) that patients, their caregivers, and clinicians can efficiently and effectively use to seek treatment recommendations or to determine eligibility for clinical trials. Patients also have the option to share their rich cancer histories (both the normalized and standardized information, as well as the raw underlying data) to advance the science of cancer. In essence, Ciitizen democratizes cancer care and research by putting data ownership and control in the hands of patients (and caregivers acting on their behalf) – the stakeholders most highly motivated to collect it, use, and share it liberally to save their own lives and the lives of others just like them.

Ciitizen provides its patient users with sole control over whether and to what extent they share the information in their Ciitizen accounts, both identifiable and de-identified. Ciitizen users will also share in any monetization of their data. In an era of decreasing trust in digital data initiatives, we believe empowering patients with decisions regarding their data is critical to helping assure that data generated from care is robustly available for research and other population health initiatives.

As of the date of submission of this proposal, Ciitizen is in beta and supporting active cancer patients on the Ciitizen platform. Ciitizen is also engaged with multiple partners, including a) supporting an observational study in non-small cell lung cancer for a large multinational life sciences partner, b) partnering with a leading advocacy organization on a rarer cancer to build a clinically validated registry, and c) working with a major health system in the Midwest to drive better clinical trial enrollment with their patient population. By the time of the AMIA annual symposium, we will have scaled up the number of patients on the platform and be able to discuss the early results of these partnerships.
Development, Design and Security Features of a REDCap -Tableau Web Data Connector (WDC) to Facilitate Operational and Clinical Self-Service Analytics
Alexander J. Tanchoco1, Olga Strachna, MS2, Stuart M. Gardos1, Jonathan Wills1, Cynthia Chu1, Joseph M. Lengfellner3, Andrew Zarski3, Ilya Sher1, Mike E. Piscitelli1, Peter D. Stetson, MD, MA2
1 Information Systems, 2 Division of Health Informatics, 3 Research and Technology Management; Memorial Sloan Kettering Cancer Center, New York, NY

Rationale, Motivation and Objectives
With growing interest for translational research in the clinical community, it’s imperative to integrate disparate sources of patient data that are captured in the electronic health record system (EHR) and augmented by data captured during the research phase1. REDCap is a popular tool within the clinical research informatics community that’s commonly used to capture clinical research study data2. While there is an existing mechanism to enable extraction of REDCap data via the REDCap repository of external modules, Tableau developers are required to request, understand the Application Programming Interface (API) mechanism and utilize the API key to authenticate to the system3. We strived to develop a human friendly mechanism for authenticating and connecting to and extracting the REDCap project data by using LDAP/AD credentials which users utilize to access enterprise Memorial Sloan Kettering (MSK) systems. The log in is linked to Datapedia, our internal MediaWiki page for collaborative sharing of knowledge about data, systems, and operational procedures. This mechanism is facilitated through a Tableau dashboard menu that shows the projects the user has access to. The user simply selects the project and does not have to remember or have a separate note of their tokens. The REDCap token is passed on to Tableau and kept in the password of the connection. This ensures that Tableau encrypts this token, just like all the embedded passwords for all other connections. Without a Web Data Connector, Tableau users who want to visualize REDCap data have to export data through CSV files which are stored on their local machines or through the network. Having the data connector facilitates improved security by eliminating data which is stored in file systems and not centrally managed. The MSK Tableau WDC also helps improve productivity and security by eliminating the tedious manual exporting and refreshing Tableau dashboards and data sources by allowing publishers to schedule the refresh automatically or by granting specific permissions for individual users to be able to perform the refreshes without having to share the credentials for connection to REDCap. This system is currently live and approved for development in production Tableau systems.

System Design Features Summary

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Description</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>General / Base</td>
<td>Enable self-service</td>
<td>Enhanced productivity and efficiency</td>
</tr>
<tr>
<td>Functionality</td>
<td>Provide a menu of authorized projects</td>
<td>For self-service, users should only have to remember their ID and password.</td>
</tr>
<tr>
<td></td>
<td>Integration with Tableau Server with the Ability to refresh data on scheduled basis when published to the Tableau Server.</td>
<td>This eliminates the manual updating of the data from REDCap enabling greater productivity by eliminating additional manual processes.</td>
</tr>
<tr>
<td>Security Requirements</td>
<td>Retain and abide by existing REDCap and Tableau Security Processes</td>
<td>Both systems are well designed and their security implementation should not be circumvented.</td>
</tr>
<tr>
<td></td>
<td>Prevent users from having to manually enter their granted API tokens.</td>
<td>Think of the REDCap API token as a user ID but without a password! It is a long string that no normal human cannot remember and they will therefore have to keep it at a place to cut and paste from. If these tokens are compromised, since REDCap is accessible on the internet, it can wreak havoc to that project.</td>
</tr>
<tr>
<td>Reporting Requirements</td>
<td>Eliminate the use of embedded passwords by using the Token as the password</td>
<td>Since the use of the token is required for the REDCap API, and because user ID is not encrypted but passwords are encrypted in Tableau, it made sense to encrypt the token which is then used to pass on to REDCap during connection. No need to keep a separate ID.</td>
</tr>
<tr>
<td></td>
<td>Integration with Active Directory</td>
<td>Provides a single user id and password for all systems.</td>
</tr>
<tr>
<td>Usability Requirements</td>
<td>Self-Service with robust documentation and integration with data definitions</td>
<td>People who are tasked with using REDCap and Tableau are likely to be highly educated and quick learners. They do need to have reference materials available. The use of Datapedia for documentation with live links to the connector allows documentation to be readily updated and tied to the rich data definitions available in Datapedia while being able to take action from within the documentation with integrated AD security.</td>
</tr>
<tr>
<td>Audit Requirements</td>
<td>Utilize both REDCap and Tableau’s security and auditing features.</td>
<td>Since the Web Data Connector is simply creatively bridging the two systems, it preserves the existing audit trails. API accesses in REDCap are logged as well as publishing and connections to the Data Sources within Tableau.</td>
</tr>
</tbody>
</table>

References
An API for LOINC release content using FHIR Terminology Services

Daniel J. Vreeman, PT, DPT, MS, FACMI1,2, Timothy Briscoe1, John Hook1, Steven Wagers1, James Agnew3, Swapna Abhyankar, MD1

1Regenstrief Institute, Inc, Indianapolis, IN, USA; 2Indiana University School of Medicine, Indianapolis, IN, USA; Smile CDR, Toronto, ON, Canada

System Purpose

LOINC® is a freely available international standard for identifying measurements, observations, and documents. Presently, LOINC has more than 80,000 users from 175 countries and has become ubiquitous in health data systems worldwide. Since its inception, Regenstrief has distributed LOINC in custom table formats and with no cost tools for linking local terms to LOINC terms. Such tools are important, but are unique to LOINC. Deploying LOINC content via a standardized application programming interface (API) could enable greater computability, facilitate use in diverse applications, and simplify access to the terminological content needed for biomedical informatics applications. The widely adopted Fast Healthcare Interoperability Resources (FHIR) specification from Health Level 7 provides a standard set of resource descriptions for health data, including terminology content. We have deployed a standardized API for LOINC that implements FHIR terminology services specifications.

System Features

FHIR defines key resources and operations in its Terminology Module, including CodeSystem, ValueSet, and ConceptMap. The CodeSystem resource specifies a set of codes drawn from one or more code systems. With input from the FHIR community, we created a canonical definition of how to represent LOINC as a CodeSystem in FHIR that extends its simple base representation to include LOINC-specific properties of each term. The LOINC CodeSystem defines several kinds of identifiers, including LOINC Codes, LOINC Parts, LOINC Answer Lists, and LOINC Answers. The ValueSet resource specifies a set of codes drawn from one or more code systems. We implemented value sets for key collections of LOINC terms, LOINC Answer Lists, and the new a LOINC Groups. The ConceptMap resource defines a mapping from a set of concepts defined in a code system to one or more concepts defined in other code systems. We implemented several sets of mappings, including from LOINC terms to IEEE device codes, LOINC terms to RadLex Playbook Procedure Codes, LOINC Parts to RadLex clinical terms, LOINC Parts to SNOMED CT terms, and more. We worked with Simpatico Intelligent Systems on extensive updates to the HAPI open source FHIR library that enabled support for these terminology resources with LOINC content, including direct import of many LOINC release artifacts and the features of the canonical definition of LOINC as a FHIR CodeSystem.

Current Use and Deployment

In September 2018, we deployed a publicly available server (accessible via free LOINC user account) powered by HAPI on Amazon Web Services. The service base URL is https://fhir.loinc.org and a detailed feature description is available at https://loinc.org/fhir. Since its launch, this service has processed an average of about 350,000 monthly requests with an average processing time of 163ms per request. We updated the terminology content with the December 2018 and June 2019 LOINC releases, and plan to do so with each upcoming LOINC release.

Conclusion

We have deployed a standardized API for LOINC content based on FHIR that provides programmatic access to current release content in a uniform and well documented format using open source software.

Funding

This work was supported by the National Center for Advancing Translational Sciences grant TR000006, the U.S. National Library of Medicine contract HHSN276201400008C, and the Office of the National Coordinator for Health Information Technology grant 90AX0021/01-00.

References

Stratification of Patient Population for enabling data-driven Clinical Interventions using I2b2

Kavishwar B. Wagholikar MBBS PhD2,3,4, Vishal P. Vernekar BE1, Akshay Zagade BE1, Yuri Ostrovsky PhD1, Shek-Wayne Chan4, Alyssa P. Goodson4, Ameet Pathak1 Corey Glynn MS3, Christopher Herrick4, Shawn N. Murphy MD PhD2,3,4

1 Persistent Systems, Pune, India and Santa Clara, CA; 2Harvard Medical School, Boston, MA; 3Massachusetts General Hospital, Boston, MA; 4Partners Healthcare Boston, MA; 5Brigham and Women’s Hospital, Boston, MA;

Introduction  An essential prelude to precision medicine is the stratification of patients into subpopulations for increased diagnostic and treatment efficacy. By efficiently matching patients to the right care management programs, improved outcomes can be obtained at less costs and in shorter time. However, electronic health records (EHR) functionality to assist clinical decision-making is restricted due to bias in health records and the difficulty in synthesizing meaningful concepts from the records. This can be resolved by using phenotyping approaches that integrate diverse data modalities. However, phenotyping methods are largely unused for driving clinical care processes, although they have been used to characterize cohorts for epidemiological studies.

Specific purpose of the system The objective of this project is to develop processes and tooling to stratify patient populations for targeted clinical care. The approach is to integrate libraries for machine learning (ML) with workflows for creating clinical annotations to train ML. The expected outcome is to identify patients for selected therapeutic pathways, which will facilitate efficiency and scalability of care delivery to the patient population.

We have developed and implemented a patient stratification process outlined in Figure 1, which incorporates feedback from ancillary clinical staff for enabling rapid development of phenotyping pipeline for identifying patients eligible for selected therapeutic pathways. The process involves (a) creation of ontologies that are geared for determining eligibility of patient for a therapy, (b) use of advanced phenotyping techniques including those driven by ML, and (c) assignment of clinical annotations for training ML algorithms.

Deployment Having deployed the process in the clinical setting for over a year, we are working on scaling the stratification process by using i2b2 plugins that we have developed. Specifically, we have developed plugins to implement key steps the patient stratification process including creation of ontologies to model and compute concepts needed at point-of-care, and incorporation of annotations from the clinical staff to provide inputs for improving the phenotyping algorithms. In this session, we will demonstrate the process for a real-world use case but using simulated patient data and demonstrate the developed i2b2 plugins.1

References

Acknowledgements
This work was supported by R00-LM011575, R01-HG009174, Partners healthcare and Persistent Systems.
ESTABLISHING AN ADAPTIVE SUSTAINABLE ENTERPRISE-LEVEL BIOMEDICAL DATA SHARING PLATFORM

Susan Tenney, Ph.D., Alexandra Shlionskaya, M.S., Michael Keller, Ph.D., Alexandra Dawson, M.S., Alissa Fujimoto, M.S., Andrijana Dabic, B.S., Rajni Samavedam, M.P.H.  
Booz Allen Hamilton, Inc.

Introduction
The scientific value of data sharing can only be fully realized if research investigators can readily access and utilize analyzable clinical and other biomedical data. Many research organizations lack the infrastructure and governance at the enterprise level to archive diverse data types, and the ability to distribute this data in accordance with patient privacy, security, and other government regulations. To address these challenges, Booz Allen Hamilton designed a data sharing platform which provides a sustainable infrastructure and streamlined mechanism for researchers to share and discover biomedical data. The platform balances the needs of researchers and their studies with the requirements mandated by institutional and federal policies which govern the sustainment and reuse of shared data.

Conceptual Data Sharing Platform Architecture
The platform utilizes scalable and secure cloud environment, flexible microservice architecture, data-driven design, and user-friendly, accessible front-end to deliver state of the art, policy-compliant, data sharing solution that is agnostic to research areas, data types, sources of the shared data. These design principles enable the secure centralization of complex and heterogenous data types generated from clinical research, in a manner that promotes FAIR (Findable, Accessible, Interoperable and Reusable) data sharing principles. To achieve consistent data representation across studies, the platform uses a harmonized metadata model to represent diverse data types in a unified, structured, and extensible way. This metadata driven approach leverages biomedical terminologies and standards to provide meaningful representation of stored data for streamlined data discovery, access, and retrieval.

Conclusion
The functionality and technical solution of this data sharing platform was designed to accommodate the growing needs of the scientific research community. The policies and governance set in place take into consideration the multi-stakeholder landscape of clinical data sharing, to ensure a transparent and streamlined data sharing process while mitigating the risks of sharing individual participant data. Collectively, this platform would serve as a valuable data sharing resource to maximize the value of existing research data by facilitating broader reuse and secondary analysis.